REFERENCE ONLY

UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author JULIOUS. S.A.

COPYRIGHT

This is a thesis accepted for a Higher Degree of the University of London. It is an unpublished typescript and the copyright is held by the author. All persons consulting the thesis must read and abide by the Copyright Declaration below.

COPYRIGHT DECLARATION

I recognise that the copyright of the above-described thesis rests with the author and that no quotation from it or information derived from it may be published without the prior written consent of the author.

LOANS

Theses may not be lent to individuals, but the Senate House Library may lend a copy to approved libraries within the United Kingdom, for consultation solely on the premises of those libraries. Application should be made to: Inter-Library Loans, Senate House Library, Senate House, Malet Street, London WC1E 7HU.

REPRODUCTION

University of London theses may not be reproduced without explicit written permission from the Senate House Library. Enquiries should be addressed to the Theses Section of the Library. Regulations concerning reproduction vary according to the date of acceptance of the thesis and are listed below as guidelines.

A. Before 1962. Permission granted only upon the prior written consent of the author. (The Senate House Library will provide addresses where possible).

B. 1962 - 1974. In many cases the author has agreed to permit copying upon completion of a Copyright Declaration.

C. 1975 - 1988. Most theses may be copied upon completion of a Copyright Declaration.

D. 1989 onwards. Most theses may be copied.

This thesis comes within category D.

☑ This copy has been deposited in the Library of UCL

☐ This copy has been deposited in the Senate House Library, Senate House, Malet Street, London WC1E 7HU.

University College London

Designing Clinical Trials with Uncertain Estimates of Variability

Steven A. Julious

A dissertation submitted to the

University of London

For the degree of

Doctor of Philosophy

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Rationale for the Thesis

One of the most important steps in the design of a clinical trial is the estimation of the sample size. For example for a superiority trial, where the data are expected to take a Normal form, the sample size (to achieve a stated power) would be based on a given clinically meaningful difference and an estimate of the population variance. This estimate of the population variance is traditionally based on the assumption of a known sampling variance when in reality this is unknown and has to be estimated. The variance estimate would be derived from an earlier similarly designed study (or a combination from several previous studies) and its precision would depend on its degrees of freedom. There is a need therefore for methods to be developed to deal with the problem of estimating sample size with imprecisely estimated variances.

Outcome of the Thesis

This thesis provides solutions for the calculation of sample sizes that allow for the imprecision of the estimates used in the calculations. It also shows how the traditional formulae give sample sizes that are too small.

The solutions given are for the calculation of sample sizes for different types of trial (superiority, non-inferiority, equivalence, bioequivalence and trials for a given precision) and different forms of data (Normal, binary and ordinal).

For Normal data a solution that uses the non-central t-distribution is given, while for binary and ordinal data numerical methods are proposed. For non-inferiority and equivalence trials with a binary outcome it is demonstrated that simple Bayesian methods add value to calculations.

Conclusions

Standard sample size calculations are shown to have limitations. The main limitation being that no account is made of the imprecision of the estimates used in the calculations. Methods are described in this dissertation that account for these limitations.

It is hoped the results would be useful to any researcher calculating a sample size for a prospective clinical study.

Acknowledgements

I'd like to thank Stephen Senn for his support as my supervisor throughout my PhD and for continuing to provide support when taking up his new role at the University of Glasgow. I'd like to express my appreciation to Rex Galbraith from University College London for taking up the reins as my primary supervisor upon Stephen leaving.

I wish to thank my previous employer GlaxoSmithKline (GSK) for encouraging me to take on a PhD and for funding my tuition fees while I did it. Thanks also to GSK for providing data for inclusion in my dissertation.

Individuals in GSK I'd like to acknowledge are Byron Jones for being my internal supervisor and Roger Owen (now of Novartis) for being an invaluable sounding board. I'd also like to thank Mike Campbell from the University of Sheffield for his comments on the final draft.

Finally, I'd like to express my gratitude to Anna Moran for sorting all the headings, tables and figures into a presentable format.

Dedication

"Knowledge must come through action; you can have no test which is not fanciful, save by clinical trial"

Adapted from Sophocles (496 BC-406 BC)

Publications

Detailed below are the publications, published in journals and given at conference, that have come from the dissertation.

Journal Publications

Julious, S.A. (2005). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* (Letter) **24:**3383-4

Julious, S.A. (2005). Issues with number needed to treat. *Statistics in Medicine* (Letter) **24:**3233-5

Julious, S.A. (2005). Sample size of twelve per group rule of thumb for a pilot study. *Journal of Pharmaceutical Statistics* **4:**287-91.

Julious, S.A. (2005). Why do we use pooled variance analysis of variance? *Journal of Pharmaceutical Statistics* **4:**3-5.

Julious, S.A. (2004). Tutorial in Biostatistics: Sample sizes for clinical trials with Normal data. *Statistics in Medicine* **23:**1921-86.

Julious, S.A. (2004). Designing Clinical Trials with Uncertain Estimates of Variability. *Journal of Pharmaceutical Statistics* **3:**261-8.

Julious, S.A. (2004). Using Confidence Intervals Around Individual Means to Assess Statistical Significance between Two Means. *Journal of Pharmaceutical Statistics* **3:**217-22.

Julious, S.A. (2004). Sample size re-determination for repeated measures studies. *Biometrics* **60:** 284-5 (Letter).

Julious, S.A. (2001). Inference and estimation in the change point regression problem. *Journal of the Royal Statistical Society, Series D* **50(1):** 51-61.

Julious, S.A. (2000). Repeated measures in clinical trials: analysis using means summary statistics and its implications for design. *Statistics in Medicine* **19:** 3133-3135 (Letter).

Julious, S.A. and Debarnot, C.A.M. (2000). Why are pharmacokinetic data summarised as arithmetic means. *Journal of Biopharmaceutical Statistics* **10(1):** 55-71.

Julious, S.A. and Owen, R.J. (2006). Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharmaceutical Statistics* **6(1):** 29-37

Julious, S.A. and Patterson, S.D. (2004). Sample sizes for estimation in clinical research. *Journal of Pharmaceutical Statistics* **3:**213-5.

Julious, S.A. and Swank, D. (2005). Moving statistics beyond the individual clinical trial- applying decision science to optimise a clinical development plan. *Journal of Pharmaceutical Statistics* **4**:37-46.

Julious, S.A., Walker, S., Campbell, M., George, S.L. and Machin, D. (2000). Determining sample sizes for cancer trials involving quality of life instruments. *British Journal of Cancer* **83(7):** 959-963.

Julious, S.A. and Zariffa, N. (2002). The ABC of pharmaceutical trial design: some basic principles. *Journal of Pharmaceutical Statistics* **1:** 45-53.

Conference Presentations

Julious, S.A. and Swank, D. (2005). Moving statistics beyond the individual clinical trial- applying simple decision science to optimise a clinical development plan. Drug Information Association, Washington.

Julious, S.A. (2004). Sample sizes for non-inferiority studies with binary data. International Society for Clinical Biostatistics, Leiden.

Julious, S.A. and Khandker, R.K. (2003). Estimating effect sizes for novel health outcomes: stroke impact scale. *DIA Meeting on Patient Outcomes:* Baltimore (Given by Khandker).

Julious SA (2002). Designing early phase trials with uncertain estimates of variability. International Society for Clinical Biostatistics Conference , Dijon.

Julious, S.A. (2002). The number needed to treat: clinically useful measure of treatment effect? Conference of Statisticians in the Pharmaceutical Industry, London

Julious SA (2001). Sample size calculations for early phase trials with uncertain estimates of variability. Conference of Statisticians in the Pharmaceutical Industry, Chester.

McClung, C. Quessy, S., Julious, S., Segretti, A. and Blum, D. (2004). Placebo response rates in geographical location in COX-2 inhibitor trials of rheumatoid arthritis (RA) and osteoarthritis (OA).

Table of Contents

List of Figures

List of Tables

1. CHAPTER 1 – INTRODUCTION

This chapter describes the background of randomised controlled clinical trials and the main factors that should be considered in their design. The description of the issues associated with clinical trial design will made in the context of a regulated pharmaceutical setting. The different types of clinical trial, for different objectives, will then be described in detail. It will be highlighted how these different objectives impact on study design with respect to derivation of formulae for sample size calculations.

The chapter will then describe, through a real example, the limitation of conventional sample size calculations and how these limitations may be assessed *a priori* when designing a trial. Finally, the chapter will describe the motivation for the PhD.

1.1. Background to Randomised Controlled Trials

Since the first 'modern' randomised clinical trial was reported [Medical Research Council, 1948], clinical trials have become a central component in the assessment of new therapies. They have contributed to improvements in healthcare as measured by an increase in life expectancy by an average of three to seven years and relief of poor quality of life related to chronic disease by an average of five years [Chalmers, 1998; Bunker, Frazier and Mosteller, 1994].

The primary objective of any clinical trial is to obtain an unbiased and reliable assessment of a given regimen response independent of any known or unknown prognostic factors i.e. they ensure that there is no systematic difference between treatments. Clinical trials are therefore designed to meet this primary objective [Julious and Zariffa, 2002]. They do this first by ensuring, as near as possible, that the patients studied in the various regimen arms are objectively similar with reference to all predetermined relevant factors other than the regimens themselves (e.g. in terms of disease severity, demography, study procedures etc). Second, by making sure that the assessment of the regimen response is independent of a given subject's regimen and finally through inclusion of an appropriate control to quantify a given regimen response. To ensure the primary objective is met Julious and Zariffa [2002] described how the essential principles of clinical trial design can be summarised in terms of the ABC of 'Allocation' at random, 'Blinding' and 'Control' group with these principals holding regardless of the type trial.

1.2. Types of Clinical Trial

When planning a trial one essential step is the calculation of a sample size as studies that are either too small or too large may be judged unethical [Altman, 1980]. For example, a study that is too large could have met the objectives of the trial before the actual study end had been reached, and so some patients may have unnecessarily entered the trial. A trial that is too small will have little chance of meeting the study objectives, and patients may be put through the potential trauma of a trial for no

tangible benefit. This chapter, based on the work of Julious [2004a], will now discuss in detail the computation of sample sizes appropriate for:

1. Superiority trials.

2. Equivalence trials.

3. Non-inferiority trials.

4. As good as or better trials.

5. Bio-equivalence trials.

6. Trials to a given precision.

A distinction therefore is drawn to emphasise differences in trials designed to demonstrate 'superiority' and trials designed to demonstrate 'equivalence' or 'non-inferiority'. This is discussed with an emphasis on how differences in the null hypothesis can impact on calculations. The ICH guidelines ICH E3 [1996] and ICH E9 [1998] provide general guidance on selecting the sample size for a clinical trial. The ICH E9 [1998] guideline states that:

> "The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed. This number is usually determined by the primary objective of the trial ....The method by which the sample size is calculated should be given in the protocol together with any quantities used in the calculations (such as variances, mean values, response rates, event rates, differences to be detected)."

This thesis is primarily written on the premise that just two treatments are to be compared in the clinical trial and two study designs will be discussed: parallel group and cross-over designs.

With a parallel group design subjects are assigned at random to the two treatments to form two treatment groups. It is hoped at the end of the trial that the two groups are the same in all respects other than the treatment received so that an unbiased assessment of treatment affect can be made.

With a cross-over trial all subjects receive both the treatments but it is the order that subjects receive the treatments which is randomised. The big assumption here is that prior to starting the second treatment all subjects return to baseline and that the order which subjects receive treatment does not affect their response to treatment. Cross-over trials cannot be used therefore in degenerative conditions, where subjects get worse over time. Also, they are more sensitive to bias than parallel group designs [Julious and Zariffa, 2002].

The assumption in the dissertation is that there is just one primary outcome to be compared and that there are not multiple endpoints. Koch and Ganksy [1996] give an

overview on the topic of multiple endpoints whilst the CPMP [2002] have issued guidelines.

1.3. Assessing Evidence from Trials

Since it is rarely possible to collect information on an entire population, the aim of clinical trials (in the context of the dissertation) is to use information from a sample to draw conclusions (or make inferences) about the population of interest. This inference is facilitated through making assumptions about the underlying distribution of the measurement such that an appropriate theoretical model can be applied to describe how the measurement in the population as a whole from the sample.

Note it is usual *a priori* to any analysis to make an assumption as to the underlying distribution of your measure. These assumptions are then to be investigated through various plots and figures for the observed data.

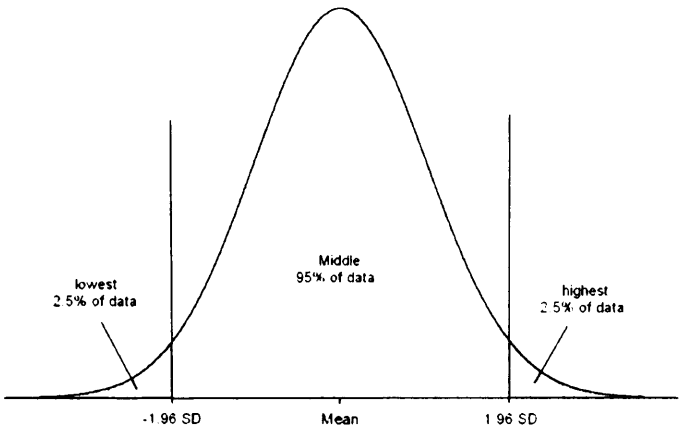
In the context of this dissertation the population is a theoretical concept used for describing an entire group. One way of describing the distribution of a measurement in a population is by use of a suitable theoretical probability distribution.

1.3.1. The Normal Distribution

The Normal, or Gaussian distribution (named in honour of C.F.Gauss, 1777-1855, German mathematician) is the most important theoretical probability distribution.

The distribution curve of data which are Normally distributed has a characteristic shape; it is bell-shaped, and symmetrical about a single peak (Figure 1.1). The Normal distribution is described completely by two parameters, the mean (μ) and the standard deviation (σ). This means that for any Normally distributed variable, once the mean and variance (σ^2) are known (or estimated), it is possible to calculate the probability distribution for that population.

Figure 1-1. The Normal distribution



1.3.2. The Central Limit Theorem

The central limit theorem (or the law of large numbers) states that given any series of independent, identically distributed random variables, their means will tend to a Normal distribution as the number of variables increases. Put another way, the distribution of sample means drawn from a population will be Normally distributed whatever the distribution of the actual data in the population as long as the samples are large enough.

Each mean estimated from a sample is an unbiased estimate of the true population mean and using the Central Limit Theorem one can infer 95% of sample means will lie within 1.96 standard deviations of the population mean. As we do not usually know the population mean the more important inference is that with the sample mean we are 95% confident that the population mean will fall within 1.96 standard deviations of the sample mean.

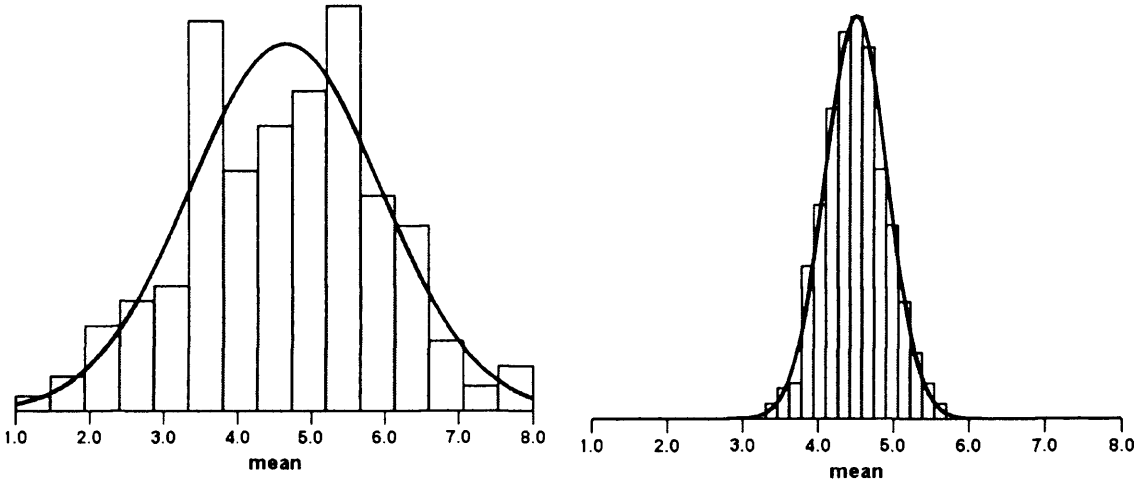
The Normal distribution and the Central Limit Theorem are important as they underpin much of the subsequent statistical theory outlined both in this and subsequent chapters. This is because although only Chapter 2 discusses calculations for clinical trials where the primary outcome is anticipated to take a Normal form, approximation to the Normal distribution (and what to do when Normal approximation is inappropriate) is important to subsequent chapters on Binary (Chapters 3 and 4) and Ordinal (Chapter 5) data.

To illustrate the Central Limit Theorem, consider the random numbers 0 to 9. The distribution of these numbers in a random numbers table would be uniform. That is to say that each number has an equal probability of being selected and the shape of the probability density function of theoretical distribution is represented by a rectangle. According to the Central Limit Theorem, if you were to select repeated random samples of the same size from this distribution, and then calculate the means of these different samples, the distribution of the means would be approximately Normal and this approximation would improve as the size of each sample increased. Figure 1.2a represents the distribution of the sample means for 500 simulated samples of size 5.

Even with such a small sample size the approximation to the Normal is remarkable, whilst repeating the experiment with samples of size 50, improves the fit to the Normal distribution (Figure 1.2b).

Figure 1-2. Distribution of means from 500 samples

a: Samples of size 5, mean=4.64, sd=1.29 **b:** Samples of size 50, mean=4.50, sd=0.41



In reality, as one usually only take a single sample, we can use the Central Limit Theorem to construct an interval within which we are reasonably confident the true population mean will be included i.e. through calculation of a confidence interval.

Application of the Central Limit Theorem is frequently made, even when non-parametric approaches are being undertaken. It is not unknown when bootstrapping (to calculate confidence interval) for the bootstrap distribution of the mean to take a Normal form. This is because for all practical purposes one is repeating the exercise described for Figure 1.2.

1.3.3. Frequentist Approaches

The descriptions of the issues associated with clinical trial design in this dissertation are made in the context of a regulated pharmaceutical setting. In this context trials are assessed through *a priori* declaring a null hypothesis, depending on the objective of the trial, and then formally testing this null hypothesis through the empirical trial data.

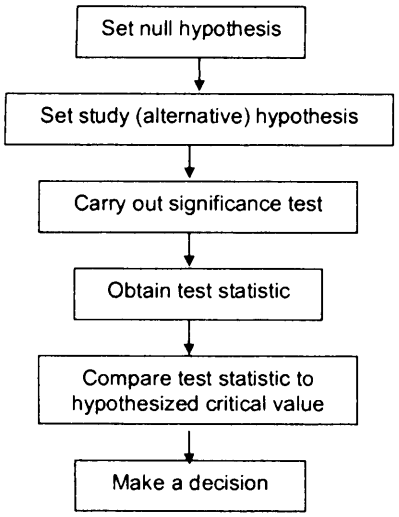
1.3.3.1. Hypothesis testing and Estimation

Consider the hypothetical example of a study designed to examine the effectiveness of two treatments for migraine. In the study patients are randomly allocated to two groups corresponding to either treatment A or treatment B. It may be that the primary objective of the trial is to investigate whether there is a difference between the two groups with respect to migraine outcome; in this case we could carry out a significance test and calculate a P-value (hypothesis testing). Alternatively it may be that the primary objective is to quantifying the difference between treatments together with a corresponding range of plausible values for the difference; in this case we would calculate the difference in migraine response for the two treatments and the associated confidence interval for this difference (estimation).

1.3.3.2. Hypothesis Testing

Figure 1.3 describes the steps in the process of hypothesis testing. At the outset it is important to have a clear research question and know what the outcome variable to be compared is. Once the research question has been stated, the null and alternative hypotheses can be formulated. The null hypothesis (H_0) usually assumes that there is no difference in the outcome of interest between the study groups. The study or alternative hypothesis (H_1) usually states that there is a difference between the study groups.

Figure 1-3. Hypothesis testing: the main steps



In lay terms the null hypothesis is what one is investigating whilst the alternative is what one wishes to show. Later in the chapter there will be a discussion of the main

types of trial objectives and their hypotheses. For example, for a superiority trial when comparing a new migraine therapy against control what one is investigating is whether there is no difference between treatments. One therefore wishes to show is that this null hypothesis is false and demonstrate that there is a difference at a given level of significance.

In general, the direction of the difference (for example: that treatment A is better than treatment B) is not specified, and this is known as a two sided (or two tailed) test. By specifying no direction one investigates both the possibility that A is better than B and the possibility that B is better than A. If a direction is specified this is referred to as a one sided test (one tailed) and one would be evaluating only whether A is better then B as possibility of B being better than A is of no interest. There will be further discussion of one tailed and two tailed tests when describing the different types of trial later in the chapter.

A common misunderstanding about the null and alternative hypotheses, is that when carrying out a statistical test, it is the alternative hypothesis (that there is a difference) that is being tested. This is not the case – what is being examined is the null hypothesis, that there is no difference between the study groups; one conducts a hypothesis test in order to establish how likely (in terms of probability) it is that we have obtained the results that we have obtained, if there truly is no difference in the population.

For the migraine trial, the research question of interest is:

> '*For patients with chronic migraines which treatment for migraine is the most effective?*'

There may be several outcomes for this study, such as the frequency of migraine attacks, the duration of individual attacks or the total duration of attacks. Assuming one is interested in reducing the frequency of attacks, then the null hypothesis, H_0, for this research question is:

> '*There is no difference in the frequency of attacks between treatment A and treatment B groups*'

and the alternative hypothesis, H_1, is:

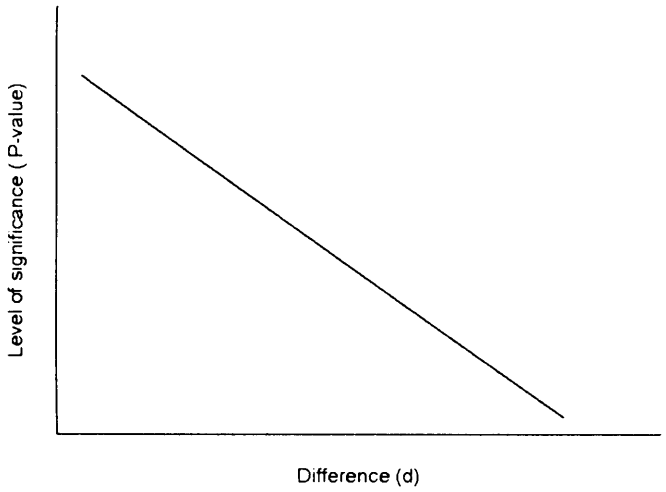
> '*There is a difference in the frequency of attacks between the two treatment groups*'.

Having set the null and alternative hypotheses the next stage is to carry out a significance test. This is done by first calculating a test statistic using the study data. This test statistic is then compared to a theoretical value under the null hypothesis in order to obtain a P-value. The final and most crucial stage of hypothesis testing is to make a decision, based upon the P-value. In order to do this it is necessary to understand first what a P-value is and what it is not, and then understand how to use it to make a decision about whether to reject or not reject the null hypothesis.

So what does a P-value mean? A P-value is the probability of obtaining the study results (or results more extreme) if the null hypothesis is true. Common misinterpretations of the P-value are that it is either the probability of the data having arisen by chance or the probability that the observed effect is not a real one. The distinction between these incorrect definitions and the true definition is the absence of the phrase when the null hypothesis is true. The omission of 'when the null hypothesis is true' leads to the incorrect belief that it is possible to evaluate the probability of the observed effect being a real one. The observed effect in the sample is genuine, but what is true in the population is not known. All that can be known with a P-value is, if there truly is no difference in the population, how likely is the result obtained (from the sample). Thus a small P-value indicates that difference we have obtained is unlikely if there genuinely was no difference in the population – it gives the probability of obtaining the study results (or results more extreme) (difference between the two study samples) if there actually is no difference in the population.

In practice, what happens in a trial is that the null hypothesis that two treatments are the same is stated i.e. A=B or A-B=0. The trial is then conducted and a particular difference d is observed where A-B=d. Due to pure randomness even if the two treatments are the same you would seldom observe A-B=0. Now if d is small (say a 1% difference in the frequency of attacks) then the probability of seeing this difference under the null hypothesis is very high say P=0.995. If a larger difference is observed then the probability of seeing this difference by chance is reduced, say d=0.05 then the P-value could be P=0.562. As the difference increases therefore so the P-value falls such that a d=0.20 may equate to a P=0.021. This relationship is very simply (i.e as linearly) illustrated in Figure 1.4 as d increases then the P-value (under the null hypothesis) fall.

Figure 1-4. Illustration of the relationship between the observed difference and the P-value under the null hypothesis



It is important to remember that a P-value is a probability and its value can vary between 0 and 1. A 'small' P-value, say close to zero, indicates that the results obtained are unlikely when the null hypothesis is true and the null hypothesis is rejected. Alternatively, if the P-value is 'large', then the results obtained are likely when the null hypothesis is true and the null hypothesis is not rejected. But how small is small? Conventionally the cut-off value or two sided significance level for declaring that a particular result is statistically significant is set at 0.05 (or 5%). Thus if the P-value is less than this value the null hypothesis (of no difference) is rejected and the result is said to be statistically significant at the 5% or 0.05 level (Table 1.1). For the example above, if the P value associated with the mean difference in the number of attacks was 0.01, as this is less than the cut-off value of 0.05 one would say that there was a statistically significant difference in the number of attacks between the two groups at the 5% level.

Table 1-1. Statistical Significance

	One say that our results are statistically significant if the P-value is less than the significance level (α), usually set at 5%	
	$P < 0.05$	$P \geq 0.05$
Result is	Statistically significant	Not statistically significant
Decide	That there is sufficient evidence to reject the null hypothesis and accept the alternative hypothesis	That there is insufficient evidence to reject the null hypothesis ↑ One cannot say that the null hypothesis is true, only that there is not enough evidence to reject it

The choice of 5% is somewhat arbitrary and though it is commonly used as a standard level for statistical significance its use is not universal. Even where it is, one study that is statistically significant at the 5% level is not usually enough to change practice; replication is required. For example to get a regulatory license for a new drug usually two statistically significant studies are required at the 5% level which equates to a single study at the 0.00125 significance level. It is for this reason that larger 'super' studies are conducted to get significance levels that would change practice i.e. a lot less than 5%.

Where the setting of a level of statistical significance at 5% comes from is not really known. Much of what one refers to as statistical inference is based on the work of R.A. Fisher (1890-1962) who first used 5% as a level of statistical significance acceptable to reject the null hypothesis. One theory is that 5% was used because Fisher published some statistical tables with different levels of statistical significance and 5% was the middle column (another is that 5 is the number of toes on Fisher's foot which maybe is just as plausible).

An exercise to do, with students say, in order to demonstrate empirically that 5% is a reasonable level for statistical significance is to toss a coin and tell the students whether a head or a tails has been observed. But keep saying heads. After around 6 tosses one asks the students when they stopped believing we were telling the truth. Usually about half would say after 4 tosses and half after 5. The probability of getting 4 heads in a row is 0.063 and the probability of getting five heads in a row is 0.031; hence 5% is a figure about which most people would intuitively start to disbelieve an hypothesis!

The significance level of 5% has to a degree become a tablet of stone. To such a degree that it is not unknown for P-values to be presented as P=0.04999999 as P must be less than 0.05 to be significant and written to 2 decimal places P=0.05 is considered to present far less evidence for rejection of the null hypothesis than P=0.04999999.

Though the decision to reject or not reject the null hypothesis may seem clear cut, it is possible that a mistake may be made, as can be seen from the shaded cells of Table 1.2. For example a 5% significance level means that we would only expect to see the observed difference (or one greater) 5% of the time under the null hypothesis. Alternatively one can rephrase this to state that even if the two treatments are the same 5% of the time we will conclude that they are not and we will make a Type I error. Therefore, whatever is decided, this decision may correctly reflect what is true in the population: the null hypothesis is rejected, when it is fact false or the null hypothesis is not rejected, when in fact it is true. Alternatively, it may not reflect what is true in the population: the null hypothesis may be rejected, when it is fact true which would lead us to a false positive and making a Type I error, (α); or the null hypothesis may not be rejected, when in fact it is false. This would lead to a false negative, and making a Type II error, (β). Acceptable levels of the Type I and Type II error rates are set before the study is conducted. As mentioned above the usual level for declaring a result to be statistically significant is set at a two sided level of 0.05 prior to an analysis i.e. the type I error rate (α) is set at 0.05 or 5%. In doing this we are stating that the maximum acceptable probability of rejecting the null when it is in fact true (committing a type 1 error, α error rate) is 0.05. The P-value that is then obtained from our analysis of the data gives us the probability of committing a Type I error (making a false positive error).

Table 1-2. : Making a decision

Decide to:	The null hypothesis is actually: False	True
Reject the null hypothesis	Correct	Type 1 Error (α)
Not reject the null hypothesis	Type 2 Error (β)	Correct

The P value. This is the probability of concluding that there is a difference, when in fact there is no difference, i.e. the probability of making a false positive mistake

This represents a well powered study – one that is able to detect a difference when there truly is a difference

The probability that a study will be able to detect a difference, of a given size, if one truly exists is called the Power of the study and is the probability of rejecting the null hypothesis when it is actually false (probability of making a Type II error, β). It is usually expressed in percentages, so for a study which has 90% power, there is a probability of 0.9 of being able to detect a difference, of a given size, if there genuinely is a difference in the population. An underpowered study is one which lacks the ability, i.e. has very low power, to detect a difference when there truly is a difference. The concepts of power and Type I and II errors will be dealt with further in a later in this chapter and throughout the dissertation as these are important components of sample size calculation.

1.3.3.3. Estimation

Statistical significance does not necessarily mean the result obtained is clinically significant or of any practical importance. A P-value will only indicate how likely the results obtained are when the null hypothesis is true. It can only be used to decide whether the results are statistically significant or not, it does not give any information about the likely size of the clinical difference. Much more information, such as whether the result is likely to be of clinical importance can be gained by calculating a confidence interval. Although earlier in the chapter discussion about the 95% confidence interval was in the context of the mean, it is possible to calculate a confidence interval for any estimated quantity (from the sample data), such as the mean, median, proportion, or even a difference. It is a measure of the precision (accuracy) with which the quantity of interest is estimated (in the case of the migraine trial, the quantity of interest is the mean difference in the number of migraine attacks).

Technically, the 95% confidence interval is the range of values within which the true population quantity would fall 95% of the time if the study were to be repeated many times. Crudely speaking, the confidence interval gives a range of plausible values for the quantity estimated; although not strictly correct it is usually interpreted as the range of values within which there is 95% certainty that the true value in the population lies.

For the migraine example the mean difference in the number of attacks between the groups, was 3 attacks per month with 95% confidence interval for this difference of 1.2 to 4.8 attacks per month. Thus, whilst the best available estimate of the mean difference was 3 attacks per month, it could be as low as 1.2 or as high as 4.8 attacks per month, with 95% certainty. As the confidence interval excludes 0 one can infer from the observed trial that it is unlikely that there is no difference between treatments. In fact as one has calculated a 95% confidence interval one can deduce that the statistical significance is less than 5%. The actual P-value associated with this difference was 0.01 and given that it is less than 5% one can conclude that the difference is statistically significant at the 5% level.

As confidence intervals are so informative and from them one can infer statistical significance as well as quantify plausible values for the population effect there is a growing consensus that only confidence intervals should be reported for studies. In this chapter also precision based trials are described where the design and analysis are based around confidence intervals. However, it is unlikely that P-values will ever be eliminated as a way to quantify differences.

1.3.3.4. Statistical and Clinical Significance

Discussion so far dealt with hypothesis testing and estimation. However, in addition to statistical significance, it is useful to consider the concept of clinical significance. Whilst a result may be statistically significant, it may not be clinically significant (relevant/important) and conversely an estimated difference that is clinically important may not be statistically significant. For example consider a large study comparing two treatments for high blood pressure; the results suggest that there is a statistically significant difference (P=0.023) in the amount by which blood pressure is lowered. This P-value relates to a difference of 3mmHg between the two treatments. Whilst the difference is statistically significant, it could be argued that a difference of 3mmHg is not clinically important. This is supported but the 95% confidence interval of 0.5 to 4.5mmHg. Hence, although there is a statistically significant difference this difference may not be sufficiently large enough to convince anyone that there is a truly important clinical difference.

This is not simply a trivial point. Often P-values alone are quoted and inferences about differences between groups are made based on this one statistic. Statistically significant P-values may be masking differences that have little clinical importance. Conversely it may be possible to have a P-value greater than the magic 5% but for there to be a genuine difference between groups: absence of evidence does not equate to evidence of absence.

The issue of clinical significance is particular important for non-inferiority and equivalence trials, discussed later in the chapter, where margins are set which confidence intervals must preclude. P-values are seldom quoted. These margins are interpreted in terms of clinically meaningful differences.

1.3.4. Bayesian Approaches

The discussion to date has been in the context of frequentist based trials where a null hypothesis is set up front; an experiment is conducted and based on the strength of the evidence observed the null hypothesis is 'accepted' or 'rejected'. The decision as to whether accept or reject is based on the P-value and the confidence intervals. This frequentist approach is the basis of all pharmaceutical regulatory trials which are the motivation for this dissertation.

The frequentist approach in some ways is somewhat naïve however. Prior to the start of the pivotal (basis for license) phase III program more than 50 trials may have been initiated or completed and yet this additional work is not included in the interpretation of the final trial(s).

In simple terms Bayesian approaches do account for this additional work (or beliefs) by setting priors before the start of a study. Once the trial has been completed the observed data are combined with the priors to form a posterior distribution for the treatment response. From this posterior distribution (95%) credibility intervals can then be calculated for the true value. This credibility interval provides a range in which there is a 95% chance the true value will lie.

First time into man studies are good examples to illustrate the utility of Bayesian methods. Table 1.3 gives the scheduled doses for an escalating dose first time into man study across 5 cohorts.

Table 1-3. Schedule doses in a hypothetical first time into man study

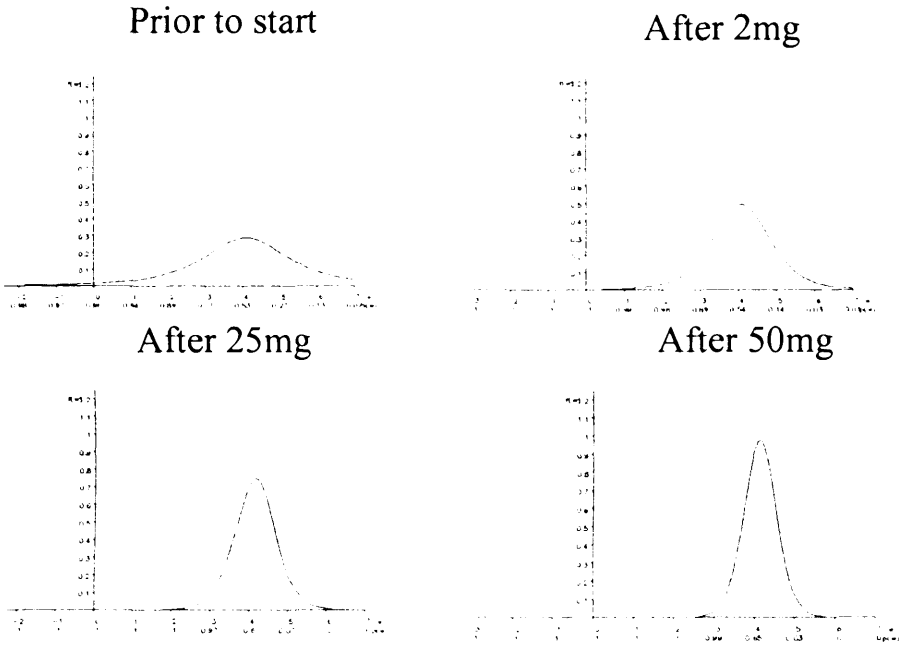
Cohort	Scheduled Doses
1	Placebo, 1mg, 2mg or 5mg
2	Placebo, 10mg, 20mg or 40mg
3	Placebo, 80mg, 150mg or 300mg
4	Placebo, 500mg, 750mg or 1,500mg
5	Placebo, 2,500mg, 5,000mg or 7,500mg

If a person naïve to drug development was asked to select a cohort to enter the study they'd probably choose cohort 1. When pushed they'd likely say because these are the lowest (and hence safest) doses. Someone familiar with drug development would choose cohort 4 or even cohort 5. When asked they'd say because the safety information from the other cohorts for what is a completely new chemical entity in man would have been received from the other cohorts by then, so with this prior knowledge a later cohort would be best. This same argument would be the retort for cohort 1 naively having the lowest doses as it is not known for certain *a priori* that these doses are the lowest as these are merely predictions. As the methodology that does these predictions, allometric scaling, is little more complex than what could be managed by a

9 year old with a ruler, it is not unknown for the first cohort to start with the maximum dose with subsequent cohort doses adjusted in light of this observed data.

Note in the context of drug development "first time into man" is not a sexist nomenclature as usually new chemical entities do use healthy males in the very early studies (except for treatments, such as hormone therapies, to be solely for women). The reason for this is that these studies will usually start up to a year before reproductive toxicity data becomes available to allow the entry of fertile female volunteers. In fact it could be argued there is an evolutionary order to drug development as rats are the tox species for men and men are the tox species for women.

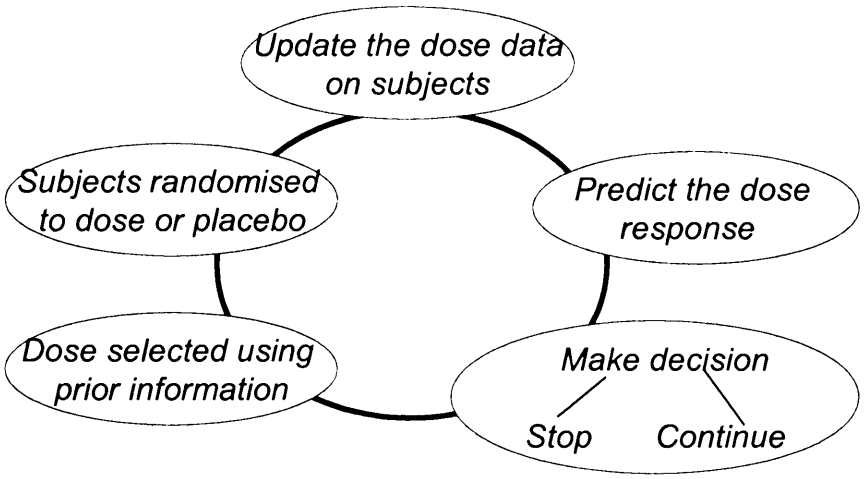
Figure 1-5. Posterior densities for log AUC at 60mg for an untested subjected



The intuitive application of Bayesian methods is highlighted in Figure 1.5 taken from Whitehead, Zhou, Patterson et al [2001]. This figure gives the posterior distribution for 60mg in a dosing cohort prior to any doses being given; and after 2mg, after 25mg and after 50mg. The prior information for the initial posterior is obtained initially from pre-clinical or other sources. As data comes in new posterior densities are derived, taken as a weighted sum of the prior and observed data. A consequence is the more the data one has the better the estimates – as evidenced by the narrowing of the posterior distribution. Figure 1.6 gives an illustration of the procedure. The Bayesian methods in this context are therefore quantifying how intuitively first time into man studies are applied.

Bayesian methods in the context of assessing toxicity in early oncology trials are routinely applied [O'Quigley, Pepe and Fisher, 1990; O'Quigley and Shen, 1996].

Figure 1-6. Assigning subjects to dose



A further example of the application of simple Bayesian methods in clinical development is given in Chapter 6 where Bayesian methods are applied in designing clinical development plans [Julious and Swank, 2005].

A final example is given Table 1.4 where Bayesian methods were applied to frame a go/no go decision [Owen, 2002]. Here the decision was in terms of the probability of achieving the safety profile based on Phase II data in Phase III. The table was used as an illustration of the number of adverse events one could see on active over those on placebo to achieve this profile.

Table 1-4. Maximum number of adverse events observed on active to ensure with 90% certainty that the difference from placebo is less than 5%

	Given Number of Adverse Events on Placebo (n=200)								
Active Sample Size	0	1	2	3	4	5	6	7	9
200	5	6	7	8	8	9	10	10	11
150	3	4	4	5	5	6	6	7	7
100	1	2	2	2	3	3	3	4	4

As a rule Bayesian methods are routinely applied throughout a drug development program except for the final quantification of proof – which would be a frequentist P-value. This is because as well as first time into man studies Bayesian methods are be applied in pharmacokinetic modelling [for example see: Judson, Peiming, Peng, Verweij, Racine and Donato di Paulo, 2005]; safety assessment [Cowell, Dawid, Hutchinson, Roden and Spiegelhalter, 1992]; application of decision science [Julious and Swank, 2005]; safety monitoring [Fayers, Ashby, and Parmar, 1997]; go/ no go decisions [Owen, 2002] and dose response assessment [for example see: Krams, Lees, Hacke, Grieve, Orgogozo and Ford, 2003;. Lunn, Wakefield and Racine-Poon, 2001]. Although there are a number of texts discussing the Bayesian design and analysis of trials [Spiegelhalter, Abrams, Myles, 2004; Spiegelhalter, Freedman and Parmer, 1995; Joseph, du Berger, and B'elisle, 1997; Pham-Gia and Turkkan, 1992] and at a basic level [Spiegelhalter, Myles, Jones, and Abrams, 1999] Bayesian methods tend to not be applied as the definitive assessment of proof in a regulatory setting. Until SAS launches a procedure PROC BAYES this situation is likely to remain the same.

The dissertation will reflect the current realities as although Bayesian methods will be applied throughout, the assumption will be that the formal assessment of proof will be through a frequentist hypothesis test.

1.4. Superiority Trials

In a superiority trial the objective is to determine whether there is evidence of a difference in the comparison of interest between the regimens with reference to the null hypothesis that the regimens are the same. The null (H_0) and alternative (H_1) hypotheses may take the form:

H_0: The two treatments have equal effect with respect to the mean response ($\mu_A = \mu_B$).

H_1: The two treatments are different with respect to the mean response ($\mu_A \neq \mu_B$).

In the definition of the null and alternative hypotheses μ_A and μ_B refer to the mean response on regimens A and B respectively. In testing the null hypothesis there are two errors one can make:

I. Rejecting H_o when it is actually true.

II. Retaining H_o when it is actually false.

As described earlier in the chapter these errors are usually referred to as Type I and Type II errors [Neyman and Pearson, 1928, 1933, 1936 and 1938]. The aim of the sample size calculation is to find the minimum sample size for a fixed probability of Type I error to achieve a value of the probability of a Type II error. The two errors are commonly referred to as the regulator's (Type I) and investigator's (Type II) risks and by convention are fixed at rates of 0.05 and 0.10 or 0.20 respectively. The Type I and Type II risks carry different weights as they reflect the impact of the errors. With a Type I error medical practice may switch to the investigative therapy with resultant costs whilst with a Type II error medical practice would remain unaltered.

In general, one usually thinks not in terms of the Type II error but in terms of the power of a trial (1-probability of a Type II error), which is the probability of rejecting the H_o when it is in fact false. Key trials should be designed to have adequate power for statistical assessment of the primary parameters. The Type I error rate is usually taken as standard for a superiority trial is 5%. The power that is becoming to be considered as standard is 90% with the minimum considered being 80%. It is debatable as to which level of power one should use although it should be noted that, compared to a study with 90% power, with just 80% power one is doubling Type II error for only a 25% saving in sample size.

As an aside it was Neyman and Pearson who introduced the concept of the two types of error, the Type I and Type II, in the 1930s. The labelling of these two types of error was arbitrary though as the authors simply listed the two types of error that could be made as sub bullets which were numbered with the prefixes of I and II. Subsequently the authors then referred to the errors as errors of Type I and errors of Type II. If these sub bullets had had different labelling, of A and B say, then statistics would have had a different nomenclature.

The purpose of the sample size calculation is hence to provide sufficient power to reject H_o when in fact some alternative hypothesis is true. For the calculation one must have a pre-specified value, for difference in the means, for the alternative hypothesis, 'd' [Campbell, Julious and Altman, 1995]. The amount d is chosen as a clinically important difference or effect size and is the main factor in determining a sample size. Reducing the effect size by half will quadruple the required sample size [Fayers and Machin, 1995]. Usually the effect size is taken from clinical judgement and/or is based on previous empirical experience in the population to be examined in the current trial.

Formally the aim is to calculate a sample size suitable for making inferences about a certain function of given model parameter, μ, $f(\mu)$ say. For data that take a Normal

form $f(\mu)$ will be $\mu_A - \mu_B$ i.e. the difference in means of two populations A and B. Now let S be a sample estimate of $f(\mu)$. Thus S is defined as the difference in the sample means. As one is assuming that the data from the clinical trial are sampled from a Normal population, then, using standard notation, S~N($f(\mu)$, Var (S)), giving

$$\frac{S - f(\mu)}{\sqrt{Var(S)}} \sim N(0,1).$$

A basic equation can now be developed in general terms from which a sample size can be estimated. Let α be the overall type I error level, with $\alpha/2$ of this type I error equally assigned to each tail of the two tailed test, and let $Z_{1-\alpha/2}$ denote the $(1-\alpha/2)100\%$ of a standard Normal distribution.

Thus, an upper 2-tailed, α -level critical region for a test of $f(\mu)$ =0 is

$$|S| > Z_{1-\alpha/2}\sqrt{Var(S)}\}.$$

For this critical region against an alternative that $f(\mu)$= d, for some chosen d, to have power (1-β)% one requires

$$d - Z_{1-\beta}\sqrt{Var(S)} = Z_{1-\alpha/2}\sqrt{Var(S)}, \qquad (1.3.1)$$

where β is the overall Type II error level and $Z_{1-\beta}$ is the 100(1-β)% point of the standard Normal distribution. Thus, in general terms for a 2-tailed, α -level test one requires

$$\text{Var(S)} = \frac{d^2}{(Z_{1-\beta} + Z_{1-\alpha/2})^2} \qquad (1.3.2)$$

where Var (S) will be unknown and depends on the sample size. Once Var(S) is written in terms of sample size, the above expressions can be solved to give the sample size.

1.4.1. Estimation of the Variance for Calculations

In this section and in subsequent sections one of the most important components in the sample size calculation is the variance estimate used in the calculations. This variance estimate is usually estimated from retrospective data sometimes from a number of studies. To assess the relative quality of the variance one should consider the following aspects of the trial from which the variance is obtained [Julious, 2004a]

1. Design: is the study design ostensibly similar to the one you are designing? On the basic level are the data from a randomised controlled trial - observational or other data may have greater variability. If you are undertaking a multi-centre trial is the variance estimated too from a similarly designed trial? Were the endpoints similar to those you plan to use – not just the actual endpoints but

were the times relative to treatment of both the outcome of interest and the baseline similar to you own? If there is likely to be missing data in the study was the same imputation method as you plan to use used?

2. Population: is the study population similar to your own? The most obvious consideration is to ask whether the demographics were the same but if the clinical trial conducted was a multi centre trial was it conducted in similar countries? Different countries may have different types of care (e.g. different concomitant medication) and so may have different trial populations. Was the same type of patient enrolled (the same mix of mild and moderate; the same season)? Was it conducted covering the same seasons (relevant for conditions such as asthma)?

3. Analysis: was the same statistical analysis undertaken? Not just the question of whether the same statistical test was used for the analysis but were the same covariates fitted into the model? Were the same summary statistics used?

The quality of the estimate of variance will obviously influence the strategy of an individual clinical trial and the question of variance estimation for sample size calculations will be returned to throughout this dissertation

1.5. Equivalence Trials

In certain cases the objective is not to demonstrate superiority but to demonstrate that two treatments have no clinically meaningful difference, i.e. they are equivalent. The null (H_1) and alternative (H_0) hypotheses may take the form:

H_0: The two treatment differences are different with respect to the mean response ($\mu_A \neq \mu_B$).

H_1: The two treatments have equal effect with respect to the mean response ($\mu_A = \mu_B$).

Usually these hypotheses are written in terms of a clinical difference, d, and become:

H_0: $\mu_A - \mu_B \leq -d$ or $\mu_A - \mu_B \geq +d$.

H_1: $-d < \mu_A - \mu_B < +d$.

The statistical tests of the null hypotheses are an example of an intersection-union test (IUT), in which the null hypothesis is expressed as a union and the alternative as an intersection. In order to conclude equivalence, one needs to reject each component of the null hypothesis.

Note that in an IUT, each component is tested at level α giving a composite test, which is also of level α [Berger and Hsu, 1996].

A common approach with equivalence trials is to test each component of the null hypothesis with a t test - called the Two One-Sided Test (TOST) procedure. In practice, this is operationally the same as constructing a (1-2α)100% confidence interval for $f(\mu)$ where equivalence is concluded provided that each end of the confidence interval falls completely within the interval $(-d,+d)$ [Jones, Jarvis, Lewis et al, 1996].

Note as each test is carried out at the α level of significance then, under the two null hypotheses, the overall chance of committing a type I error is less than α [Senn, 1997, 2001]. Hence, the TOST, and (1-2α)100% confidence interval, approach, is conservative. There are enhancements that can be applied but they are of no practical importance for formally powered clinical trials [Senn, 1997, 2001]. As a consequence the TOST approach will only be discussed for equivalence trials (and bioequivelence trials later). Figure 1.7 highlights how equivalence can be demonstrated through confidence intervals.

Figure 1-7. An illustration of average equivalence between two populations

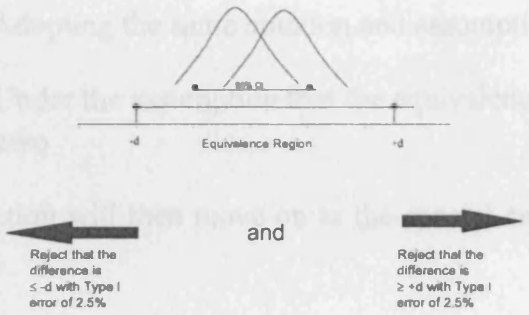
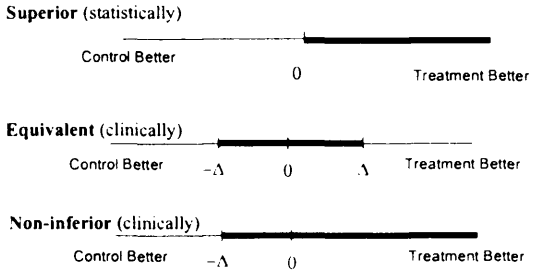


Figure 1.8 shows how confidence intervals are used to test the different hypotheses in superiority and equivalence trials. The special case of bioequivalence is covered in Section 1.8.

Figure 1-8. An illustration of difference between equivalence, non-Inferiority and superiority



* Δ is variable

ICH E10 [2000] goes into some detail in the description of equivalence trials, and the related non-inferiority trials (discussed in Section 1.5) whilst ICH E9 [1998] and ICH E3 [1996] discuss the appropriate analysis of such trials.

In this section the sample size formulae will initially be derived

I. For the general case of inequality between treatments (i.e. $f(\mu) = \Delta$)

II. Adopting the same notation and assumptions as in Section 1.4

III. Under the assumption that the equivalence bounds $-d$ and d are symmetric about zero

This section will then move on to the special case of no treatment difference replacing (i) with:

I. For the special case of no mean difference (i.e. $f(\mu)=0$).

1.5.1. General Case

As with Section 1.3 one requires

$$\frac{S - f(\mu)}{\sqrt{Var(S)}} \sim N(0,1).$$

Hence, the $(1-2\alpha)100\%$ confidence limits for a non-zero mean difference would be

$$S - \Delta \pm Z_{1-\alpha} \sqrt{Var\ S}\ .$$

To declare equivalence the lower and upper confidence limit should be within ±d

$$S - \Delta - Z_{1-\alpha}\sqrt{Var(S)} > -d \text{ and } S - \Delta + Z_{1-\alpha}\sqrt{Var(S)} < d. \quad (1.4.1)$$

Thus, for the two one sided test procedure (TOST) with this critical region there are two opportunities under the alternative hypothesis to have a Type II error for some chosen d and power (1-β)

$$\Delta + d - Z_{1-\beta_1}\sqrt{Var(S)} = Z_{1-\alpha}\sqrt{\text{Var(S)}} \text{ and } \Delta - d - Z_{1-\beta_2}\sqrt{\text{Var(S)}} = -Z_{1-\alpha}\sqrt{\text{Var(S)}} \quad (1.4.2)$$

where β_1 and β_2 are the probability of a Type II error associated with each one sided test from the TOST procedure and $\beta = \beta_1 + \beta_2$. Hence, one requires

$$Z_{1-\beta_1} = \frac{\Delta + d}{\sqrt{Var(S)}} + Z_{1-\alpha} \text{ and } Z_{1-\beta_2} = \frac{\Delta - d}{\sqrt{Var(S)}} - Z_{1-\alpha}. \quad (1.4.3)$$

Alternatively, Senn (1997) considers the calculation of the Type II error in terms of the power and hence has a slightly different nomenclature. However, they are equivalent.

1.5.2. Special Case of No Treatment Difference

With symmetric equivalence bounds one requires

$$S \pm Z_{1-\alpha}\sqrt{Var\ S},$$

Thus, to declare equivalence one should have

$$S - Z_{1-\alpha}\sqrt{Var(S)} > -d \text{ and } S + Z_{1-\alpha}\sqrt{Var(S)} < d.$$

With the TOST procedure the Type II error for some chosen d and power (1-β) will come from

$$d - Z_{1-\beta}\sqrt{Var(S)} = Z_{1-\alpha}\sqrt{\text{Var(S)}} \text{ and } -d - Z_{1-\beta}\sqrt{\text{Var(S)}} = -Z_{1-\alpha}\sqrt{\text{Var(S)}}.$$

Hence,

$$Z_{1-\beta\ 2} = \frac{d}{\sqrt{Var(S)}} - Z_{1-\alpha},$$

giving

$$Var\ (S) = \frac{d^2}{(Z_{1-\alpha} + Z_{1-\beta\ 2})^2}. \quad (1.4.4)$$

1.5.3. Choice of Type I Error and Equivalence Limit

1.5.3.1. Choice of Type I Error

Strictly speaking when undertaking two simultaneous one tailed tests setting α=0.05 would maintain an overall Type I error rate of 5%. However, the choice of the Type I error is a controversial issue. The convention for equivalence trials is to set the Type I error rate at half of that which would be employed for a two sided test used in a superiority trial i.e. α=0.025. That is, giving a Type I error rate of 2.5% [ICH E9, 1998]. However, setting the Type I error rate for equivalence trials at half that for superiority trials could be considered consistent. This is because although in a superiority trial one has a two sided 5% significance level in practice for most trials in effect what one has is a one sided investigation with a 2.5% level of significance. The reason for this is that one usually has an investigative therapy and a control therapy and it is only statistical superiority of the investigative therapy that is of interest.

Through the rest of the sections on equivalence and non-inferiority trials the assumption will be that α=0.025 and that 95% confidence intervals will be used in the final statistical analysis - although remember as discussed in this chapter in section 1.4 that the type I error is actually a little less than 2.5%.

The issue of setting an appropriate Type I error level will be discussed again in the section on bioequivalence later in this chapter.

1.5.3.2. Choice of Equivalence Limit

The discussion on setting equivalence limits in this section can also be generalised to non-inferiority trials discussed in the proceeding section. As with the choice of the Type I error the setting of the non-inferiority/equivalence limit is a controversial issue. The limit is defined as the largest difference that is clinically acceptable such that a larger difference than this would matter in clinical practice [CPMP, 2000]. This difference also cannot be "greater than the smallest effect size that the active (control) drug would be reliably expected to have compared with placebo in the setting of the planned trial" [ICH E10, 2000].

However, beyond this there is not much formal guidance. Jones, Jarvis, Lewis et al [1996] have recommended that the choice of limit be set at half the expected clinically meaningful difference between the active control and placebo. There is no hard regulatory guidance although the CPMP [1999] in a concept paper originally stated that for non-mortality studies it might be acceptable to have an equivalence limit "of one half or one third of the established superiority of the comparator to placebo, especially if the new agent has safety or compliance advantages". Although in the draft notes for guidance that followed the CPMP [2004] have moved away from such firm guidance and state, "Historically, it has been common to select as delta (a) proportion of the difference between comparator and placebo. Such an approach does not necessarily ensure superiority over placebo and there is no clinical rationale to support it". The

CPMP now talk of having a margin that ensures that there is "no important loss of efficacy" caused through switching from reference to test and that the margin could be defined from a "survey of practitioners on the range of differences that they consider to be unimportant".

Generally, the definition of the acceptable level of equivalence or non-inferiority is made with reference to some retrospective superiority comparison to placebo [Hung, Wang, Lawrence et al, 2003; D'Agostino, Massaro and Sullivan, 2003; Wiens, 2002]. Methodologies for indirect comparisons to placebo have been discussed in detail by Hasselblad and Kong [2001]. In this context the definition of the non-inferiority and equivalence limits should address the following steps [Wiens, 2002; D'Agostino, Massaro and Sullivan, 2003; Julious, 2004a].

1. One must be confident that the active control would have been different from placebo had one been employed.

2. One should be able to determine that there is no clinically meaningful difference between investigative treatment and the control.

3. Through comparing the investigative treatment to control one should indirectly be able to determine that it is superior to placebo.

Steps 1. and 3. are important as there is a view that non-inferiority and equivalence trials reward "failed" studies i.e. if one conducted a poor trial where it would not have been possible to demonstrate the active control to be superior to placebo then a poor investigative therapy may slip through, through comparison to this control. However, Julious and Zariffa [2002] point out that this may not be the case as poor studies are poor for most objectives due to their higher statistical variability.

In summary therefore one can infer that the clinical difference used for the limits of equivalence and non-inferiority will be smaller than the difference used for placebo controlled superiority trials. There also is no generic definition for its setting – its definition will need to be defined on a study by study basis with consultation with the appropriate agencies and experts.

1.6. Non-Inferiority Trials

In certain cases the objective of a trial is not to demonstrate that two treatments are different or that they are equivalent but rather to demonstrate that a given treatment is not clinically inferior compared to another i.e. that a treatment is non-inferior to another. The null (H_0) and alternative (H_1) hypotheses may take the form:

H_0: A given treatment is inferior with respect to the mean response.

H_1: The given treatment is non-inferior with respect to the mean response.

As with equivalence trials these hypotheses are written in terms of a clinical difference, d, which again equates to the largest difference that is clinically acceptable [CPMP, 2000]:

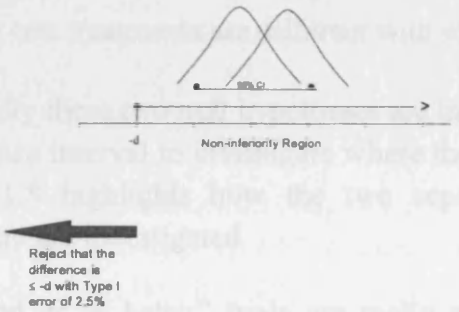
H_0: $\mu_A - \mu_B \leq -d$.

H_1: $\mu_A - \mu_B > -d$.

ICH E3 [1996] and ICH E9 [1998] go into detail on the analysis of non-inferiority trials whilst ICH E10 [2000] goes into detail as to the definition of d.

In order to conclude non-inferiority, one needs to reject the null hypothesis. In terms of the equivalence hypotheses in Section 1.5 this is equivalent to testing just one of the two components of the TOST procedure and reduces to a simple one-sided hypothesis test. In practice, this is operationally the same as constructing a (1-2α)100% confidence interval and concluding non-inferiority provided that the lower end of this confidence interval is above –d. Figure 1.9 highlights how non-inferiority can be demonstrated through confidence intervals and Figure 1.8 shows how confidence intervals are used to test the different hypotheses in superiority, equivalence and non-inferiority trials.

Figure 1-9. An illustration of average non-inferiority between two populations



Adopting the same notation and under the same assumptions as in Section 1.5 but with $f(\mu) = -\Delta$ and the additional assumption that the non-inferiority bound is *-d,* the lower $(1-2\alpha)100\%$ confidence limit is

$$S - \Delta - Z_{1-\alpha}\sqrt{Var\ S}. \qquad (1.5.1)$$

To declare non-inferiority the lower end of the confidence interval should lie above –d

$$S - \Delta - Z_{1-\alpha}\sqrt{Var\,(S)} > -d. \qquad (1.5.2)$$

For this critical region one therefore requires a $(1-\beta)100\%$ chance that the lower limit lies above $-d$. Hence,

$$Z_{1-\beta} = \frac{-d+\Delta}{\sqrt{Var(S)}} - Z_{1-\alpha} \ , \qquad (1.5.3)$$

giving

$$Var\,(S) = \frac{(d-\Delta)^2}{(Z_{1-\alpha}+Z_{1-\beta})^2} . \qquad (1.5.4)$$

1.7. As Good as or Better Trials

For certain clinical trials the objective is to demonstrate either that a given treatment is not clinically inferior or that it is clinically superior when compared to the control i.e. that the treatment is "as good as or better" than the control. Therefore two null and alternative hypotheses are being investigated in such trials. First the non-inferiority null and alternative hypotheses:

H_0: A given treatment is inferior with respect to the mean response ($\mu_A \leq \mu_B$).

H_1: The given treatment is non-inferior with respect to the mean response ($\mu_A > \mu_B$).

If this null hypothesis is rejected then a second null hypothesis is investigated:

H_0: The two treatments have equal effect with respect to the mean response ($\mu_A = \mu_B$).

H_1: The two treatments are different with respect to the mean response ($\mu_A \neq \mu_B$).

Practically these two null hypotheses are investigated through the construction of a 95% confidence interval to investigate where the lower (or upper as appropriate) bound lies. Figure 1.8 highlights how the two separate hypotheses for superiority and non-inferiority are investigated.

"As good as or better" trials are really a sub-category of either superiority or non-inferiority trials. However, in this dissertation this class of trials are put into a separate section to highlight how as good as or better trials combine the null hypotheses of superiority and non-inferiority trials into one closed testing procedure whilst maintaining the overall Type I error [Morikawa and Yoshida, 1995; Bauer and Kieser, 1996; Julious, 2004a].

To introduce the closed testing procedure this section will first describe the situation where a one-sided test of non-inferiority is followed by a one-sided test of superiority. The more general case where a one sided test of non-inferiority is followed by a two sided test of superiority is then described.

In describing "as good as or better" trials this thesis draws heavily on the work of Morikawa and Yoshida [1995]. The CPMP [2000] have recently issued a points to consider document.

1.7.1. A Test of Non-Inferiority and One Sided Test of Superiority

The null ($H1_0$) and alternative ($H1_1$) hypotheses for a non-inferiority trial can be written as:

$H1_0$: $\mu_A - \mu_B \leq -d$.

$H1_1$: $\mu_A - \mu_B > -d$.

Which alternatively can be written as

$H1_0$: $\mu_A - \mu_B + d \leq 0$.

$H1_1$: $\mu_A - \mu_B + d > 0$.

Whilst the corresponding null ($H2_0$) and alternative ($H2_1$) hypotheses for a superiority trial can be written as

$H2_0$: $\mu_A - \mu_B \leq 0$.

$H2_1$: $\mu_A - \mu_B > 0$.

What is clear from the definitions of these hypotheses is that if $H2_0$ is rejected at the α level then $H1_0$ would also be rejected. Also, if $H1_0$ is not rejected at the α level then $H2_0$ would also not be rejected. This is because $\mu_A - \mu_B + d \geq \mu_A - \mu_B$. Hence, both $H1_0$ and $H2_0$ are rejected if they are both statistically significant; neither $H1_0$ and $H2_0$ are rejected if $H1_0$ is not significant; only $H1_0$ is rejected if only $H1_0$ is significant.

Based on these properties a closed test procedure can be applied to investigate both non-inferiority and superiority whilst maintaining the overall Type I error rate without α adjustment. To do this the intersection hypothesis $H2_0 \cap H1_0$ is first investigated which, if rejected, is followed by a test of $H1_0$ and $H2_0$. In this instance $H2_0 \cap H1_0 = H1_0$ and so both non-inferiority and superiority can be investigated through the following two steps [Morikawa and Yoshida, 1995].

First investigate the non-inferiority through the hypothesis $H1_0$. If $H1_0$ is rejected then $H2_0$ can be tested. If $H1_0$ is not rejected then the investigative treatment is inferior to the control treatment.

If $H2_0$ is then rejected in the next step one can conclude that the investigative treatment is superior to the control. Else if $H2_0$ is not rejected then non-inferiority should be concluded.

1.7.2. A Test of Non-Inferiority and Two Sided Test of Superiority

The null ($H3_o$) and alternative ($H3_1$) hypotheses for a two-sided test of superiority can be written as:

$H3_o$: $\mu_A = \mu_B$.

$H3_1$: $\mu_A < \mu_B$ or $\mu_A > \mu_B$.

Which is equivalent to two one-sided tests at the $\alpha/2$ level of significance – summing to give an overall type I error of α - with the investigation of $H2_o$ against the alternative of $H2_1$ and the following hypotheses:

$H4_o$: $\mu_A \geq \mu_B$.

$H4_1$: $\mu_A < \mu_B$.

In applying the closed test procedure in this instance it is apparent that the intersection hypothesis $H1_0 \cap H3_0$ is always rejected as it is empty and so both $H1_o$ and $H3_o$ can be tested. Due to there being no intersection the following steps can be applied steps [Morikawa and Yoshida, 1995]:

1. If the observed treatment difference is greater than zero and $H3_o$ is rejected then $H1_o$ is also rejected and one can conclude that the investigative treatment is superior to control.

2. If the observed treatment difference is less than zero and $H3_o$ is rejected and $H1_o$ is not then the control is statistically superior to the investigative treatment. If $H1_o$ is also rejected then the investigative drug is worse than the control but is not inferior (practically though this may be difficult to claim).

3. If $H3_o$ is not rejected but $H1_o$ is, then the investigative drug is non-inferior compared to the control.

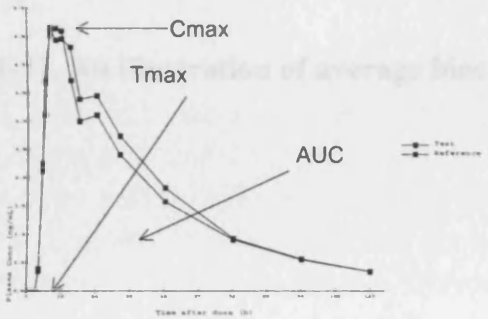
4. If neither $H1_o$ nor $H3_o$ are rejected then one must conclude that the investigative treatment is inferior to control.

Note that when investigating the $H1_o$ and $H3_o$ hypotheses, using the procedure described above, $H3_o$ will be tested at a two sided α level of significance whilst $H1_o$ will be tested at a one-sided $\alpha/2$ level of significance. Thus, the overall level of significance is maintained at α.

1.8. Assessment of Bioequivalence

Earlier in the chapter trials were described where one wished to demonstrate that the two therapies are clinically equivalent. In equivalence trials the comparators may be completely different, in terms of route of administration or even actual drug therapies, but what one wishes to determine is whether they are clinically the same. However, in bioequivalence trials the comparators are ostensibly the same - one may have simply moved manufacturing sites or had a formulation changed for marketing purposes. Bioequivalence studies are therefore conducted to show these two formulations of the drug have similar bioavailability – the amount of drug in the bloodstream. The assumption in bioequivalence trials is that if the two formulations have equivalent bioavailability then one can infer that they have equivalent therapeutic effect for both efficacy and safety. The pharmacokinetic bioavailability is therefore a surrogate for the clinical endpoints. As such one would expect the concentration time profiles for the test and reference formulations to be super-imposable, see Figure 1.10 for an example, and the two formulations to be clinically equivalent.

Figure 1-10. An example of pharmacokinetic profiles for test and reference formulations



In bioequivalence studies, therefore, one can determine whether *in vivo* the two formulations are bioequivalent by assessing whether the concentration time profiles for the test and reference formulations are super-imposable [Senn, 1998]. Assessing if the rate and extent of absorption are the same usually does this. The pharmacokinetic parameter AUC (area under the concentration curve) is used to assess the extent of absorption and the parameter Cmax (maximum concentration) is used to assess the rate of absorption. Figure 1.10 gives a pictorial representation of these parameters. If the two formulations are bioequivalent then they can be switched without reference to further clinical investigation and can be considered inter-changeable.

The null and alternative hypotheses are similar to those for equivalence studies:

H_0: The test and reference formulations give different drug exposures $(\mu_T \neq \mu_R)$.

H_1: The test and reference formulations give equivalent drug exposure $(\mu_T = \mu_R)$.

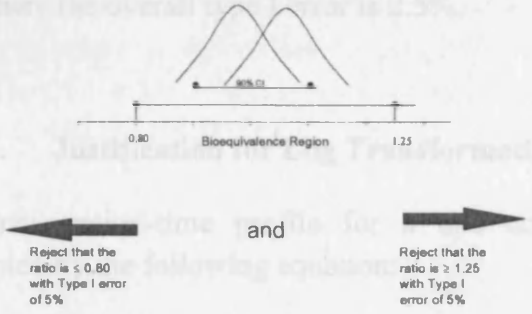
Similarly to other types of trials the objective of a bioequivalence study is to test the null hypothesis to see if the alternative is true. The 'standard' bioequivalence criteria demonstrate that average drug exposure on the test is within 20% of the reference on the log scale [FDA 2000, 2001; CPMP, 1998]. Thus, the null and alternative hypotheses can be rewritten as:

H_o: $\mu_T/\mu_R \leq 0.80$ or $\mu_T/\mu_R \geq 1.25$.

H_1: $0.80 < \mu_T/\mu_R < 1.25$.

One can declare two comparator formulations to be bioequivalent if one can demonstrate that the mean ratio is wholly contained within 0.80 to 1.25. To test the null hypothesis one undertakes two one-sided tests at the 5% level to determine whether $\mu_T/\mu_R \leq 0.80$ or $\mu_T/\mu_R \geq 1.25$. If neither of these tests hold then one can accept the alternative hypothesis of $0.80 < \mu_T/\mu_R < 1.25$. As one is performing two simultaneous tests on the null hypothesis, both of which must be rejected to accept the alternative hypothesis, the type I error is maintained at 5%. Similar to equivalence trials discussed earlier in this chapter the convention is to represent the two one-sided tests as a 90% confidence interval around the mean ratio of μ_T/μ_R which summarises the results of two one-tailed tests. Figure 1.11 highlights how average bioequivalence between two formulations can be demonstrated through 90% confidence intervals.

Figure 1-11. An illustration of average bioequivalence between two formations



In summary, a test formulation of a drug is said to be bioequivalent to its reference formulation if the 90% confidence interval for the ratio test:reference is wholly contained within the range 0.80 to 1.25, for both AUC and Cmax. As both AUC and Cmax must be equivalent to declare bioequivalence there is no need to allow for multiple comparisons.

Note this example raises the issue of loss of power when one has multiple endpoints. Here both AUC and Cmax needed to hold to declare bioequivalence and so the type I error is not inflated. However, such "and" comparisons may affect the type II error, depending on the correlation between the endpoints, as there is twice the chance to make a type II error which can impact on the power. [Koch and Ganksy, 1996; CPMP, 2002]. The most extreme situation would be for two independent "and" comparisons where the type II error is doubled. However, here AUC and Cmax are highly correlated and as one selects the highest variance from the two to calculate the sample size this means that any increase in the type II error could be offset by the fact that for either of AUC or Cmax the power is greater than 90% for the calculated sample size.

For compounds with certain indications other parameters, such as Cmin (defined as the minimum concentration over a given period) or Tmic (defined as time above a minimum inhibitory concentration over a given period), may also need to be assessed.

Note, the criteria for acceptance of bioequivalence may vary depending on factors such as which regulatory authority's guidelines are being followed and the therapeutic window of the compound being formulated and so the 'standard' criteria may not always be appropriate.

The methodology described in this section can also be applied to other types of *in vivo* assessment such as the assessment of a food [FDA, 1997]; drug interaction [CPMP 1997, FDA 1999] or special populations [FDA 1998, 1999]. The criteria for acceptance for other types of *in vivo* assessment may vary depending on the guidelines [FDA 1999] or *a priori* clinical assessment [CPMP 1997, FDA 1997, 1999].

It may be worth noting the statistical difference between testing for equivalence and bioequivalence with reference to investigating the null hypothesis. In equivalence trials the convention is to undertake two one-sided tests at the 2.5% level, which in turn are represented by a 95% confidence interval; in a bioequivalence trial two one-sided tests at the 5% level are undertaken, which are represented by a 90% confidence interval. Thus, in bioequivalence trials the overall type I error is 5% - twice that of equivalence trials where the overall type I error is 2.5%.

1.8.1. Justification for Log Transformation

The concentration-time profile for a one compartment intravenous dose can be represented by the following equation:

$c(t) = Ae^{(-\lambda t)}$,

where t is time, A is the concentration at t=0 and λ is the elimination rate constant [Julious and Debarnot, 2000]. It is evident from this equation that drug concentration in the body falls exponentially at a constant rate λ. A test and reference formulation are super-imposable, therefore only when $c_T(t) = c_R(t)$. On the log scale this is equivalent to $\log(A_T) - \lambda_T = \log(A_R) - \lambda_R$ which for $\lambda_T = \lambda_R$ (which a priori one would expect)

becomes $\log(A_T) = \log(A_R)$. Thus, on the log scale the difference between two curves can be summarised on an additive scale and indeed it is upon this scale that such pharmacokinetic parameters as the rate constant, λ, and the half life, are derived [Julious and Debarnot, 2000]. This simple rationale also follows through for statistics used to measure exposure (AUC) and absorption (Cmax) as well as the pharmacokinetic variance estimates [Lacey, Keene, Pritchard et al, 1997; Julious and Debarnot, 2000]. Hence, unless there is evidence to indicate otherwise, the data are assumed to follow a log-Normal distribution and hence the default is to analyse $\log_e$ AUC and $\log_e$ Cmax. The differences on the $\log_e$ scale (test-reference) are then back-transformed to obtain a ratio. It is the back transformed ratio and its corresponding 90% confidence interval which is used to assess bioequivalence.

1.8.2. Rationale for Using Coefficients of Variation

All statistical inference for bio-equivalence trials is undertaken on the log scale and back transformed to the original scale for interpretation. Thus, the within-subject estimate of variability on the log scale is used both for inference and sample size estimation. With the interpretation of the mean effect on the original scale it is good also to have a measure of variability also on the original scale. A measure of variability usually is the Coefficient of Variability (CV) as for log-Normally distributed data the following exact relationship between the CV on the arithmetic scale and the standard deviation, σ, on the log scale holds [Dilletti, Hauschke and Steimijans, 1991; Julious and Debarnot, 2000]

$$CV = \sqrt{\left(e^{\sigma^2} - 1\right)} \ .$$

For small estimates of σ^2 [$\sigma < 0.30$] the CV can be approximated by

$$CV = \sigma \ .$$

which relies on the first two terms of the Taylor series expansion of $\exp(\sigma^2)$. Thus, both the measure of effect and its variability can both be interpreted on the original scale. The derivation of this is result is based on the following relationships for the log-Normal distribution [Julious and Debarnot, 2000]

$$m = e^{(\mu + \sigma^2/2)},$$

$$s^2 = \left(e^{(2\mu+\sigma^2)}\right)\left(e^{\sigma^2} - 1\right),$$

where μ and σ^2 relate to the mean and variance respectively on the log-transformed scale and m and s the corresponding mean and variance on the non-transformed scales and hence

$$CV = \frac{s}{m} = \frac{\sqrt{\left(e^{(2\mu+\sigma^2)}\right)\left(e^{\sigma^2} - 1\right)}}{e^{(\mu+\sigma^2/2)}} = \sqrt{\left(e^{\sigma^2} - 1\right)} \ .$$

1.8.3. Individual and Population Bioequivalence

The assessment of bioequivalence as defined in this chapter is based on average bioequivalence in which only the formulation means are required to be equivalent. New paradigms for bioequivalence based on population and individual bioequivalence have also be been proposed [Schall and Williams, 1996; Hauck and Anderson, 1992] for which there are regulatory guidelines [FDA, 2001]. These alternative approaches also involve formulation variabilities as well as their means in the assessment bioequivalence. To calculate a sample size recommendations have been made based on simulations [FDA, 2001].

The merits of the concepts of individual and population bioequivalence are debatable and some authors have questioned the concepts [Senn, 2001]. There are a number of reasons for this. For first is that for two formulations A and B; in a study it could be possible to declare A to have individual or population bioequivalence with B while the converse is not true.

The second reason is that there is no hierarchy to the assessments. If in a study individual bioequivelence was declared between two formulations it is not then possible to automatically be enable to conclude population bioequivalence and average bioequivalence. In fact it is possible to be able to conclude individual bioequivalence and yet have a point estimate outside of the standard average bioequivalence bounds of (0.80, 1.25).

The final reason is turning the arguments for individual and population bioequivalence assessment around. The justification for their use is that they allow for an assessment of switchability and prescribability of two formulations which have greater clinical meaning. This may apply if the study conducted is in a patient population with clinical endpoints. However, bioequivalence studies are conducted in healthy volunteers using surrogate endpoints (pharmacokinetics) and so the argument pertaining to "switchability" and "prescribability" fail.

1.9. Estimation to a Given Precision

In the previous sections of the chapter calculations have been described with reference to some clinical objectives such as the demonstration of equivalence. However, often a preliminary or pilot investigation is conducted where the objective is to provide evidence of what the potential range of values is with view to doing a later definitive study [Wood and Lambert. 1999; Day, 1988; Julious and Patterson, 2004; Julious 2004a]. Such studies may also have sample sizes based more on feasibility than formal consideration [Julious, 2005d].

In given drug's development, it may be the case that reasonably reliable estimates of between-subject and of within-subject variation for the endpoint of interest in the

reference population are available, but the desired magnitude in the treatment difference of interest will be unknown. This may be the case, for example, when considering the impact of an experimental treatment on biomarkers [Biomarkers Definitions Working Group, 2001] or other measures not known to be directly indicative of clinical outcome but potentially indicative of pharmacological mechanism of action. In this situation, drug and biomarker development will be in such an early stage that no pre-specified treatment difference will be of interest nor either will statistical testing of any observed treatment difference. In such exploratory or 'learning' studies [Sheiner, 1997], what is proposed in this dissertation is that the sample size be selected in order to provide a given level of precision in the study findings, not to power in the traditional fashion for a (in truth unknown) desirable and pre-specified difference of interest.

For such studies, rather than testing a hypothesis, it is more informative to give an interval estimate or confidence interval for the unknown $f(\mu)$.

Recall that $(1-\alpha)100\%$ confidence interval for $f(\mu)$ has half-width

$$w = Z_{\alpha/2}\sqrt{Var(S)}. \qquad (1.8.1)$$

Hence, if one is able to specify a requirement for w and write Var (S) in terms of 'n' then the above expression can be solved for n. It should be noted though that if the sample size is based on precision calculations, then the protocol should clearly state this as the basis for the size of the study.

A similar situation occurs when the sample size is determined primarily by practical considerations. In this case one may quote the precision of the estimates obtained based on the half-width of the confidence interval, and provide this information in the discussion of the sample size. Again it must be clearly stated in the protocol that the size of the study was determined based on practical, and not formal, considerations.

The estimation approach also could be useful where one wishes to quantify a possible effect across several doses, or to power on a primary endpoint overall but also to have sufficient precision in given subgroup comparisons. The former of these may be a neglected consideration for clinical trials even though there is some regulatory encouragement as the CPMP [2002] Points to Consider on Multiplicity Issues in Clinical Trials states:

> "Sometimes a study is not powered sufficiently for the aim to identify a single effective and safe dose but is successful only at demonstrating an overall positive correlation of the clinical effect with increasing dose. This is already a valuable achievement. Estimates and confidence intervals are then used in an exploratory manner for the planning of future studies."

Indeed in early trial for the same sample size as doing a larger study where a single dose is powered against placebo one could undertake a well-designed study based on the precision approach with several doses estimated against placebo.

In an alternative use for the exploratory use of confidence intervals Julious [Julious, 2004c] highlights how non-overlapping 84% confidence intervals around individual means equates to a significance level of 5% for the difference between means. This result could be used if one has several doses across time points to explore possible differences between groups (not accounting for multiplicity).

1.10. Conventional Calculations and Their Limitations

Subsequent chapters in this thesis will in turn go through the conventional sample size calculations for each of the different types of clinical trial described in the previous sections of this chapter for Normal, binary and ordered categorical data.

One potential issue with conventional calculations is that they all usually rely on retrospective data to quantify the variance to be used in the calculations. Even for binary data a retrospective control prevalence will feed into a variance estimate. If this variance is therefore estimated imprecisely then it would impact on the calculations. Calculations at the moment do not allow for this imprecision.

To highlight the issues around using imprecise variance estimates in the following sub-section a worked example will be given, which was introduced by Julious [Julious, 2004b]. In the scheme of things this example is quite an important one as it is the genesis for most the consequent work that has formed the basis of this thesis.

The issues raised by this one study led to work on how to prospectively consider the sensitivity of the sample size to the assumptions about the variance (which will be discussed in detail throughout this dissertation) and to work on adaptive designs (including sample size re-estimation), a brief discussion of which will be given in Chapter 6.

The conclusion to this work are proposed new methodologies developed within this dissertation for the calculation of sample sizes, which takes account of the imprecision of the variance estimate.

1.10.1. Worked Example

A bioequivalence study was conducted to compare a test and reference formulation. Such bioequivalence studies are conducted to show that two formulations of the drug have similar bioavailability. As discussed in Section 1.7 the assumption in bioequivalence trials is that if two formulations have equivalent bioavailability then one can infer that they have equivalent therapeutic effect, for both efficacy and safety.

For the worked example the 'standard' bioequivalence criteria were used *a priori* such that bioequivalence was to be declared if the average drug exposure on the test (μ_T) was within 20% of the reference (μ_R) on the log scale [FDA, 2000, 2001; CPMP,

1998]. Thus, *a priori* it was determined that bioequivalence could be concluded if the 90% confidence interval for μ_T/μ_R is completely contained within (0.80, 1.25).

The planned study design was a two period cross-over trial (AB/BA) with AUC and Cmax being used to assess bioequivalence. As both AUC and Cmax must both be bioequivalent to declare bioequivalence there was no issue with multiplicity (although as discussed earlier when discribing bioequivalence studies there may be implications for the type II error). An estimate of the within-subject variability was available from previous studies, CVw=30%, and the mean true ratio of μ_T/μ_R was assumed to be unity. The total sample size was calculated to be 39 subjects or 20 per sequence (AB or BA). In the trial 48 subjects were recruited to ensure 40 completed (to allow for drop outs). The study was completed and the results are presented in Table 1.5

Table 1-5. Results of the example bioequivalence study

	N	Ratio	90% C.I.	CVw%
AUC	45	1.10	(0.94, 1.29)	47
Cmax	47	1.05	(0.92, 1.21)	41

From these results bio-equivalence it seems can be declared for Cmax but not for AUC. Thus, as both AUC and Cmax must hold to be able to declare bioequivalence it was concluded that the study had failed and the two formulations were not bioequivalent. However, there was some evidence that the two formulations were bioequivalent - the point estimates for the mean ratios of both AUC and Cmax were within 0.80 to 1.25.

One factor that seemed to have caused the problems was the unexpectedly high variances. Within subject CVw's of 47% and 41% were observed for AUC and Cmax respectively compared to 30% used in the sample size calculations. There was one marked outlier in the analysis but even after excluding this outlier the CVw's were 42% and 38% for AUC and Cmax respectively. There was no reason to exclude this subject from the analysis and the final analysis was presented including all subjects.

As one might imagine these results caused a great deal of contemplation upon their reporting. Especially as no study is an island for, as well as the cost of a failed study, there was an impact in timelines on the development of the asset, which were dependent upon the results of this study. Thus, there was a double impact of having to conduct another (far larger) study and also having to wait for these study results.

1.10.2. Sensitivity Analysis

In the study given in the worked example two studies, given in Table 1.6, were used to obtain a variability estimate.

Table 1-6. Within subject coefficients of variability with their corresponding degrees of freedom (DF) observed in two previous studies prior to the study undertaken in the worked example for the primary endpoints of AUC and Cmax

	AUC		Cmax	
	CVw(%)	df	CVw(%)	Df
Study 1	33%	13	20%	13
Study 2	24%	15	23%	15
Pooled	29%	28	27%	28

The maximum pooled estimate of variance, as measured by the CV, observed from these two studies was 29% for AUC which was rounded up to 30% for calculations. What was not considered in the calculations at all was the fact that this estimate of the CV was made with just 28 degrees of freedom. The assumption in the calculations therefore was that the variance used in the calculations was the population variance. However, what was not undertaken *a priori* was any sensitivity analysis of the study design to the assumptions around the sample variance. On this issue ICH E9 [1998] makes the following comment, where the emphasis is that of the author:

> "The method by which the sample size is calculated should be given in the protocol, together with the estimates of any quantities used in the calculations (such as variances, mean values, response rates, event rates, difference to be detected)..... It is important to investigate the ***sensitivity*** of the sample size estimate to a variety of deviations from these assumptions....."

The investigation of the sensitivity of the trial design to the assumptions about the variance is relatively straightforward to investigate and could have been done using the degrees of freedom of the variance estimate used in the calculations [Julious, 2004b]. First of all, one needs to calculate the sample size conventionally using an appropriate variance estimate. Next, using the degrees of freedom for this variance and the chi-squared distribution, one can calculate the upper one tailed 95th percentile for the variance

From these pooled estimates and corresponding degrees of freedom Table 1.7 could have been constructed and the maximum plausible estimate for the CV, taken as the 95th percentile, would be 38%. If this CV for the AUC was observed and not the 30% used in the calculations then the study would still have had 70% power. Thus, the study was reasonably robust to deviations about the variance estimate.

Table 1-7. Sensitivity analysis about the coefficients of variability (%) observed in two previous studies

	CVw	95th	Power for 95th
AUC	29%	38	70%
Cmax	27%	35	76%

There are a couple of points worth noting though. First, the actual CV observed in the study was greater than that estimated as the 95th percentile from the sensitivity analysis. Indeed the variance estimates for Cmax and AUC, excluding the outlying subject, fell on the 99th percentile based on the previous studies. This fact demonstrates that, although it would be nice to be wise after the event, no one has a crystal ball and that, by definition, one will always encounter unexpected variances.

The second point to highlight is that although the planned study was robust to deviations in the assumptions about the variance it was hit with the double whammy of an unexpected large variance and an unexpectedly large regimen difference – as the study was designed assuming no difference between treatments (a ratio of 1.00) when a ratios of 1.10 (for AUC) and 1.05 (for Cmax) were observed.

1.10.3. Calculating the Sample Size Accounting for the Imprecision in the Variance Estimate

As highlighted in the previous section, the pooled within-subject CVw used in the sample size calculations was 30% estimated with 28 degrees of freedom. Chapter 2 describes for bioequivalence studies the methodology for sample size calculation that would account for the fact that the variance used in the sample size calculation was estimated with just 28 degrees of freedom. Accounting for this imprecision would equate to a total sample size estimate of 44 subjects or 22 per sequence (AB or BA). To account for dropouts 54 subjects in total would need to be recruited to ensure 44 completed. Thus, accounting for the imprecision in the variance would increase the sample size by 10%. If this proportional increase in the sample size had been applied then one may have expected 48 and 52 subjects to be evaluable for the comparisons of AUC and Cmax respectively. Thus, with a 10% proportional increase in the sample size the 90% CI for AUC would have become 0.94 to 1.28 and Cmax would have become 0.92 to 1.20.

1.10.4. Moving Beyond the Conventional Calculations - Motivation for Further Work

The worked example given in the previous sub-section was both a good and bad one. Good in that it highlights the issues associated with imprecise variability estimates and the subsequent costs if this imprecision is ignored – as with conventional calculations. Bad in that it does not allow one to be wise after the event and say "if only....". However, the worked example does highlight the motivation for the work in this PhD as it demonstrates how an imprecisely estimated variance used in a sample size calculation can have a substantial impact on a study – particularly if this imprecision is not allowed for in the sample size calculation.

The issue of imprecise variability estimates is one that is most associated with early phase clinical trials, where, by definition, there is little clinical experience of a compound in man with comparatively few subjects available to provide estimates of the variability for the first formally powered study. It should be highlighted though that the

work could be generalised to later phase trials. For example, where a novel endpoint is being used for the first time within a company and the information on variability may be limited. Alternatively, when one is undertaking a later phase trial in a sub population where there may be little variability data available for the specific sub population.

In addition the logic of the issues of imprecise variability estimates for Normally distributed data can be extended to other forms of data such as that of binary (Chapters 3 and 4) and ordinal (Chapter 5) where the respective variability estimates would come from the anticipated control responses based on retrospective data for the binary or ordinal outcome of interest. The solution is not so straightforward for binary or ordinal data, however, the thesis will also discuss the solutions to these particular problems also.

Finally the thesis will finish with an investigation of issues generic to all forms of trial independent of the type of data (Chapter 6). Included in this chapter will be an investigation into issues such as adaptive designs, clinical development plans and computer intensive methods with respect to their impact on sample size calculation.

2. CHAPTER 2 - SAMPLE SIZE CALCULATIONS FOR CLINICAL TRIALS WITH NORMAL DATA

2.1. Introduction

This chapter describes the calculations for clinical trials where the expectation is that the data will take a plausibly Normal form. This chapter will discuss the standard sample size calculations for trials where the objective is to determine: superiority; equivalence; non-inferiority; estimation to a given precision and bioequivalence. These calculations will be described for both cross-over and parallel group designed studies. Much of the description of the standard calculations comes from Julious [Julious, 2004a]. This chapter will then describe how to undertake sensitivity analyses around the sample size calculations when designing the trial. Finally it will introduce calculations for the estimation of the sample size, which account for the degrees of freedom of the sample variance used in the calculations.

2.2. Aims of the Chapter

The main issues to be covered in this chapter are:

- To describe the standard sample size calculations for data anticipated to take a Normal form and how the different clinical trial objectives impact on these calculations.
- To highlight the limitation of standard calculations with respect to the assumptions about the sample variance used in the calculations.
- To introduce methodologies for undertaking sensitivity analyses, with respect to the sample variance.
- To derive a methodology for sample size calculation that takes account of the imprecision in the sample variance.
- To discuss the impact of design factors, such as covariates or repeated measures, on sample size calculations.
- To describe how Bayesian methods could be applied to explore and account for the assumptions in the sample size calculations.

2.3. Superiority Trials

2.3.1. Parallel Group Trials

2.3.1.1. Sample Sizes Estimated Assuming the Population Variance to be Known

As discussed in Chapter 1, in general terms for a 2-tailed, α -level test one requires

$$\text{Var (S)}=\frac{d^2}{(Z_{1-\beta}+Z_{1-\alpha/2})^2}, \tag{2.2.1}$$

$$Var\ (S)=\frac{\sigma^2}{n_A}+\frac{\sigma^2}{n_B}=\frac{r+1}{r}.\frac{\sigma^2}{n_A}, \tag{2.2.2}$$

where σ^2 is the population variance estimate and $n_B = rn_A$. Note: (2.2.2) is minimised when r = 1 for fixed n. Substituting (2.2.2) into (2.2.1) gives the standard sample size result which does not allow for the imprecision about the variance [Brush, 1998; Lemeshow, Hosmer, Klar et al, 1990]

$$n_A=\frac{(r+1)(Z_{1-\beta}+Z_{1-\alpha/2})^2\sigma^2}{rd^2}. \tag{2.2.3}$$

Note that in this section, and throughout the chapter for parallel group trials with Normal data, the assumption will be made that the variances in each group are equal i.e. that $\sigma_A^2=\sigma_B^2=\sigma^2$. This assumption is referred to as homoscedasticity. There are alternative formulae for the case of unequal variances [Schouten, 1999; Singer, 2001] and Julious [2005a] has described how the assumptions of homogeneity impacts on statistical analysis.

When the clinical trial has been conducted and the data have been collected and cleaned for analysis it is usually the case that for the analysis the population variance, σ^2, is considered unknown and a sample variance estimate, s^2, is used instead of σ^2. As a consequence of this a t-statistic as opposed to a Z-statistic is used for inference. This fact should be represented in the sample size calculation rewriting (2.2.3) so that t- as opposed to Z-values are used. Hence, if the population variance is considered unknown for the analysis the following could be used

$$n_A \geq \frac{(r+1)(Z_{1-\beta}+t_{1-\alpha/2,n_A(r+1)-2})^2\sigma^2}{rd^2}. \tag{2.2.4}$$

As n_A appears on both the left and right side of (2.2.4) it is best to rewrite the equation in terms of power and then use an iterative procedure to solve for n_A

$$1-\beta=\Phi\left(\sqrt{\frac{rn_Ad^2}{(r+1)\sigma^2}}-t_{1-\alpha/2,n_A(r+1)-2}\right), \tag{2.2.5}$$

where $\Phi(\bullet)$ is the defined as the cumulative density function of N(0,1). However, it is not just a simple case of replacing Z values with t values for the case of when a sample variance is being used in the analysis. In this situation the power should be estimated from a cumulative t-distribution as opposed to a cumulative Normal [Senn, 1993; Brush, 1988; Chow, Shao and Wang, 2002; Julious 2004a]. The reason for this is that by replacing σ^2 with s^2 (2.2.5) becomes

$$1-\beta = P\left(\sqrt{\frac{rn_A d^2}{(r+1)s^2}} - t_{1-\alpha\ 2,n_A(r+1)-2}\right),$$

where $P(\bullet)$ denotes a cumulative distribution defined below. This equation can in turn be rewritten as

$$1-\beta = P\left(\frac{\sqrt{rn_A}d/\sqrt{(r+1)}\sigma}{\sqrt{s^2/\sigma^2}} - t_{1-\alpha\ 2,n_A(r+1)-2}\right),$$

by dividing top and bottom by σ^2. Thus, one has a Normal over a square root of an independent chi-squared (divided by its degrees of freedom), which is a t-distribution. More specifically, in fact, as the power is estimated under the alternative hypothesis, and that under this hypothesis $d \neq 0$, the power should hence be estimated from a non-central t distribution with degrees of freedom $n_A(r+1)-2$ and non-centrality parameter $\sqrt{rn_A/(r+1)\sigma^2}$ [Senn, 1993; Brush, 1988; Chow, Shao and Wang, 2002; Kupper and Hafner, 1989; Julious 2004a]. Thus, (2.2.5) can be rewritten as

$$1-\beta = 1-\text{probt}\left(t_{1-\alpha\ 2,n_A(r+1)-2},\ n_A(r+1)-2,\ \sqrt{\frac{rn_A d^2}{(r+1)\sigma^2}}\right), \qquad (2.2.6)$$

where $\text{probt}\left(\bullet,\ n_A(r+1)-2,\ \sqrt{rn_A d^2/(r+1)\sigma^2}\right)$ denotes the cumulative distribution function of a Student's non-central t distribution with $n_A(r+1)-2$ degrees of freedom and non-centrality parameter $\sqrt{rn_A d^2/(r+1)\sigma^2}$ Note here, the notation, $\text{probt}\left(\bullet,\ n_A(r+1)-2,\ \sqrt{rn_A d^2/(r+1)\sigma^2}\right)$, is the same as that used in the statistical package SAS notation. Note also that when d=0 then one has a standard t distribution.

Practically one could use (2.2.3) for the initial sample size calculation and then calculate the power for this sample size using (2.2.6), iterating as necessary to the required power is reached. To further aid in these calculations a correction factor of $Z_{1-\alpha/2}/4$ can be added to (2.2.3) to allow for the Normal approximation [Guenther, 1981; Campbell, Julious and Altman, 1995; Julious 2004a]

$$n_A = \frac{(r+1)\left(Z_{1-\beta} + Z_{1-\alpha\ 2}\right)^2 \sigma^2}{rd^2} + \frac{Z^2_{1-\alpha\ 2}}{4}. \qquad (2.2.7)$$

For quick calculations the following formula to calculate sample sizes, with 90% power and a two-sided 5% type I error rate, can be used

$$n_A = \frac{10.5\sigma^2}{d^2}\frac{(r+1)}{r}, \qquad (2.2.8)$$

or for r=1

$$n_A = \frac{21\sigma^2}{d^2}.$$

Equations (2.2.7) and (2.2.8) are close solutions to (2.2.6), giving estimates only one or two lower and thus provide quite good initial estimates. Equation (2.2.5) is closer to (2.2.6) mainly giving the same result and occasionally underestimating by just 1. Table 2.1 gives sample sizes using (2.2.6) for various standardised differences ($\delta = d/\sigma$).

Table 2-1. Sample sizes for one group, n_A (n_B=n_A) in a parallel group study for different standardised differences and allocation ratios for 90% power and a two sided type I error of 5%

	Allocation ratios			
δ	1	2	3	4
0.10	2103	1577	1402	1314
0.20	527	395	351	329
0.30	235	176	157	147
0.40	133	100	89	83
0.50	86	64	57	53
0.60	60	45	40	37
0.70	44	33	30	28
0.80	34	26	23	21
0.90	27	21	18	17
1.00	23	17	15	14

2.3.1.2. Worked Example

An investigator wishes to design a hypertension trial with equal allocation between groups where the clinical effect of interest is a reduction in blood pressure compared to control of 4mmHg (d). The expected standard deviation in the population in which the trial is to be undertaken is 20mmgHg (σ). Thus, the standardised difference equates to $\delta = d/\sigma = 4/20 = 0.20$. For the Type I and Type II errors fixed at 5% and 10% respectively, (2.2.8) gives 526. Using this sample size to initiate iterations in (2.2.6) one gets the following steps:

Iteration	n	Power
1	526	0.8993
2	527	0.9004

Thus, the sample size required is 527 subjects in each arm of the trial and a total sample size of 1054. Alternatively one could look up the standardised effect of 0.20 in Table 2.2, which gives the same sample size.

Table 2-2. Sample sizes for one group, n_A (n_B=n_A) in a parallel group study for different standardised differences with 90% power and a two sided type I error of 5% along with the power corresponding to the 95th percentile of the variance for difference degrees of freedom

		Degrees of Freedom										
δ	n	5	10	15	20	25	30	40	50	75	100	200
0.10	2103	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.84
0.20	527	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.85
0.30	235	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.85
0.40	133	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.85
0.50	86	0.34	0.53	0.62	0.67	0.70	0.73	0.76	0.78	0.80	0.82	0.85
0.60	60	0.34	0.53	0.62	0.67	0.70	0.73	0.76	0.78	0.80	0.82	0.85
0.70	44	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.85
0.80	34	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.85
0.90	27	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.770	0.80	0.82	0.85
1.00	23	0.35	0.55	0.64	0.69	0.72	0.74	0.77	0.79	0.82	0.83	0.86

2.3.1.3. Sensitivity Analysis about the Variance Used in the Sample Size Calculations

The sensitivity of the trial design to the variance is relatively straightforward to investigate and can be done using the degrees of freedom of the variance estimate used in the calculations. This concept was described by Julious [2004b]. First of all, one needs to calculate the sample size conventionally using an appropriate variance estimate. Next, using the degrees of freedom for this variance and the chi-squared distribution, one can calculate the upper one tailed 95th percentile, say, for the variance using the following formula

$$s_p^2(95) < \frac{df}{\chi^2_{0.95,df}} s_p^2. \qquad (2.2.9)$$

Then this upper estimate of the variance can be used in (2.2.6), to investigate the power. This would give an assessment of the sensitivity of the study to deviations from the variability assumptions, by investigating a study's power to an extreme plausible value that the variance could take.

Table 2.2 gives the sample size per group required for different standardised differences for a parallel group superiority trial along with the sensitivity of these sample sizes to the 95th percentile of the variance for different degrees of freedom. From Table 2.2 it seems that to ensure 50% and 80% power with the 95th percentile about the variance one would require 10 and 75 degrees of freedom respectively.

2.3.1.4. Worked Example

The investigator as in the earlier worked example wishes investigate the sensitivity of a hypertension trial with respect to the variance used in the calculations. Recall the mean reduction of interest was 4mmHg (d) that, with an expected standard deviation in the

population of 20mmgHg (σ), gives a standardised difference of 0.20 and hence a target sample 527 subjects in each arm. However, the estimate of the population variance used in the calculation was only from a relatively small pilot investigation, which had just 10 degrees of freedom for the sample variance. From (2.2.9) one can calculate an upper one tailed estimated of the variance – and consequent standard deviation – from the 95th percentile. This estimate could be taken as a high plausible value for the variance. With such few degrees of freedom an upper estimate of the standard deviation is relatively high at 31.85mmHg. Putting this standard deviation into (2.2.6), with the sample size fixed at 527, one would expect to have power of 53% if the standard deviation observed was nearer to 31.85mmHg than the 20mmHg used in the original calculations.

Alternatively one can use Table 2.2. With 10 degrees of freedom for the variance one can see that a high plausible variance estimate would give one 53% power - the same as calculations in the previous paragraph.

2.3.1.5. Optimising the Variance Estimates

What is apparent from what has been described in this section so far is that the more degrees of freedom about the variance one has the less sensitive calculations are to assumptions about the variance. This is as one would expect. However, often when calculating a sample size a great deal of information is thrown away. One approach that is common when calculating sample sizes is to tabulate all the variances estimated from previously observed studies and then take the maximum of these variances. Another is to calculate a crude arithmetic mean across the studies to obtain an overall estimate. Both these approaches may be appropriate if the variances originate from studies of similar size; however, in many instances the studies are of diverse sample size with diverse estimates of the variance. The most extreme variance estimates are also often those from the smallest studies and thus by taking the maximum or by taking the arithmetic mean one may be giving undue weight to the studies with the poorest estimate of the variance.

If there are several studies with variance estimates available then it is recommended that an overall estimate of the population variance is obtained from the following formula [Julious 2004b]

$$s_p^2 = \frac{\sum_{i=1}^{k} df_i s_i^2}{\sum_{i=1}^{k} df_i}, \qquad (2.2.10)$$

where k is the number of studies, s_i^2 is the variance estimate from the i th study and df_i is the degrees of freedom about this variance. The pooled variance estimate, s_p^2 is the minimum variance unbiased estimate of the population variance and is estimated with the following degrees of freedom

$$df_p = \sum_{i=1}^{n} df_i \,. \qquad (2.2.11)$$

This estimate of the variance has a number of obvious advantages. The main advantage is that appropriate weight is given to the variances with the smallest and largest degrees of freedom. Another advantage is that by combining the variance estimates one is maximising the degrees of freedom about the overall estimate.

Note the assumption in these calculations is that the variances are all estimates of the sample population variance. If there is true heterogeneity between variances then different would need to be considered. The issue of heteroscadicity of trials will be discussed in Chapter 6.

2.3.1.6. Calculations Taking Accounting of the Imprecision of the Variance Used in the Sample Size Calculations

Sample size determination is now reconsidered in the context of Normally distributed observations with common variance σ^2. However, as previously highlighted, typically σ^2 would be unknown but the choice of sample size, which depends crucially on σ^2, has to be decided before any observations in the prospective trial have been made. This impasse is overcome with an assumption about σ^2. The simplest approach is just to assume that σ^2 is known and takes an 'assumed value' which is the basis for traditional sample size formulae discussed previously in this chapter. In reality the 'assumed value' is obtained from an estimate s^2 of σ^2 from previous similarly designed studies using the same endpoint.

This section addresses the issue that the traditional sample size formulae do not take account of the uncertainty about σ^2. It is assumed that the estimate s^2 of σ^2 is of the 'standard' type so that the ratio $(ms^2)/\sigma^2$ would have a chi-squared distribution with a known number of degrees of freedom, m. In particular this implies that s^2 is an unbiased estimate of σ^2. The most common situation is revisited where the sample size must be decided before any data are observed in the planned trial. The context of this problem was introduced by Julious [2004b] and the methodology described by Julious and Owen [Julious, 2002a; Julious and Owen, 2006]

The sample size and power calculations are based on an estimate of σ^2 but once the prospective study is complete any tests of hypotheses would be based entirely on the 'current' estimate of variance from the study data alone.

The context considered throughout this section, as throughout the chapter, is that of a two arm parallel group study with a given fixed ratio r for the numbers of patients in the two arms, n patients in one arm and r×n patients in the other arm so there is a total of $N=(1+r)n_A$ patients. Choice of n_A for a parallel study is made according to the criteria explained above for testing H_0 versus H_1. In this context the best linear unbiased estimate $\hat{\theta}$ of θ would be adopted,

$$\hat{\theta} = \bar{x}_A - \bar{x}_B . \qquad (2.2.12)$$

which has variance $(r+1)\sigma^2/(rn_A)$ so $\hat{\theta}$ is distributed $Normal(\theta, (r+1)\sigma^2/(rn_A))$. When the prospective trial is run and analysed there will be a standard estimator $\hat{\sigma}^2$ of σ^2 and hence the following ratio follows a Student's t distribution on $n_A(r+1)-2$ degrees of freedom,

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}\sqrt{(r+1)/(rn_A)}} \sim t_{n_A(r+1)-2} . \qquad (2.2.13)$$

Hence for a given type I error level, α, a 2-tail α -level critical region for a test of the null hypothesis, $\theta = 0$, would be given by

$$\left|\hat{\theta}\right| > t_{1-\alpha\ 2,n_A(r+1)-2}\hat{\sigma}\sqrt{(r+1)/(rn_A)} , \qquad (2.2.14)$$

where, for $0 < P < 1$, $t_{P,\nu}$ denotes the value such that $prob(T \leq t_{P,\nu}) = P$ where T has a standard Student's t–distribution on ν degrees of freedom. However, at the time of the power calculation $\hat{\sigma}^2$ is of course unknown and hence the boundaries of the critical region (2.2.14) are unknown. Traditionally the determination of n_A is made on the basis of the following argument. If σ^2 were known then (2.2.14) would become

$$\left|\hat{\theta}\right| > t_{1-\alpha\ 2,n_A(r+1)-2}\sigma\sqrt{(r+1)/(rn_A)} , \qquad (2.2.15)$$

and if $\theta = d$ then for given σ this would occur with probability

$$\Psi(d) = \Phi\left(\sqrt{\frac{rn_A d^2}{(r+1)\sigma^2}} - t_{1-\alpha\ 2,n_A(r+1)-2}\right),$$

where Φ denotes the cumulative distribution function (c.d.f.) of a standard Normal distribution. Let us call $\Psi(d)$ the 'true' power of the study for given n_A and σ^2. Hence, the power for given n_A and σ^2 is given by

$$\Psi(d) = \Phi\left(\sqrt{\frac{rn_A d^2}{(r+1)\sigma^2}} - t_{1-\alpha\ 2,n_A(r+1)-2}\right). \qquad (2.2.16)$$

Here (2.2.16) is equivalent to (2.2.6) discussed earlier in the chapter. This is a simpler result than (2.2.6) although at a practical level they give the same sample size (occasionally underestimating by 1). The implication of using (2.2.16) will be discussed later. Note that here and throughout the following arguments d, α, β and r are fixed with the power dependent on n_A and σ.

In the spirit of the 'usual' argument, since σ^2 is in fact unknown it is replaced by s^2 in (2.2.16) and then this expression is set greater than or equal to the desired power $1-\beta$. This results in the 'traditional' result, given earlier in this chapter, for (the integer) n_A to achieve a power of at least $1 - \beta$

$$n_A \geq \frac{(r+1)\left(Z_{1-\beta} + t_{1-\alpha\ 2,n_A(r+1)-2}\right)^2 s^2}{rd^2}. \qquad (2.2.17)$$

As noted previously, however, that as n_A appears on both sides of this inequality it is obtained by iteration and a good starting value for this iteration can be obtained by replacing $t_{1-\alpha/2,n_A(r+1)-2}$ in (2.2.17) by $Z_{1-\alpha/2}$, the corresponding percentile for the standard Normal distribution, so n_A would be chosen to be the least integer exceeding n_A' where

$$n_A' = \frac{(r+1)\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 s^2}{rd^2} \qquad (2.2.18)$$

For the moment ignoring the constraint of n_A to integer values, if n would satisfy (2.2.17) exactly then the true power of the study would be (by substitution into (2.2.16))

$$\Psi(d) = \Phi\left(\sqrt{\frac{rd^2}{(r+1)\sigma^2}\frac{(r+1)s^2(t_{1-\alpha/2,n_A(r+1)-2} + Z_{1-\beta})^2}{rd^2}} - t_{1-\alpha/2,n_A(r+1)-2}\right)$$

hence,

$$\Psi(d) = \Phi\left(\frac{s(t_{1-\alpha/2,n_A(r+1)-2} + Z_{1-\beta})}{\sigma} - t_{1-\alpha/2,n_A(r+1)-2}\right). \qquad (2.2.19)$$

Here (2.2.19) gives power as a function of d, n_A, r, s, σ α and β. $\Psi(d)$ is now something slightly different giving the power obtained when n_A is chosen in order to achieve a power using an estimate of the standard deviation in place of the population value. It is logically different from (2.2.16) as n_A is forced to satisfy (2.2.17) from which (2.2.19) is derived.

Note that d does appear in the right hand side of (2.2.19) although of course n_A depends implicitly on d though (2.2.17) and (2.2.16) from which (2.2.19) is derived. In fact of interest now is the anticipated power for a given sample size (n_A) for fixed d, r, α and β.

Note also that in the special case where σ^2 is assumed known, $s^2 = \sigma^2$ (which can be thought of as corresponding to $m = \infty$) and it can be seen that in this case the right hand side of (2.2.19) reduces to $1-\beta$ exactly.

In general however the power $\psi(d)$ is a random variable because s^2 is a random variable. This in turn implies that n_A and hence $t_{1-\alpha\ 2,n_A(r+1)-2}$ are random variables too. In order to assess the way this affects the choice of n an expression for the *expected* power is required

$$\mathbf{E}(\Psi) = \mathbf{E}\{\mathbf{E}(\Psi \mid t_{1-\alpha\ 2,n_A(r+1)-2})\}, \qquad (2.2.20)$$

where the outer expectation is with respect to $t_{1-\alpha/2,n_A(r+1)-2}$. Now, if $Y \sim \sqrt{\chi_m^2/m}$ then for any constants a and b then the expectation of (2.2.19) is [Ellison, 1964]

$$E[\Phi(bY-a)] = \text{probt}(b,m,a). \qquad (2.2.21)$$

Similar to earlier in the chapter here probt$(\bullet,m,a)$ denotes the cumulative distribution function of a Student's non-central t distribution with m degrees of freedom and non-centrality parameter a. To prove the result given in (2.2.21) some notation is first introduced. Let R and W denote random variables, then $E(R|W)$ is defined to be the random variable h(W) where h(w) ≡ E(R |W= w). Also E{E(R |W)} is defined to be E[h(W)]. Now, to evaluate the left hand side of (2.2.21) the problem can be thought of as a special case of trying to determine

$$E[\Phi(bY-Z)], \qquad (2.2.22)$$

where $Y = \sqrt{\chi^2_m / m}$ and $Z \sim N(\xi,\theta^2)$ independently of Y and b can take any value for some ξ and θ^2 to be chosen. Note that later in this section Y will be set as $Y = s/\sigma$. It can be shown that for two random variables R and W

$$E[R]=E\{E[R|W]\}, \qquad (2.2.23)$$

provided only that these expectations exist [Rao, 1965]. What one wishes to establish now is the result

$$E[\Phi(bY-Z)] = P(X \le bY - Z), \qquad (2.2.24)$$

where X is a standard Normal independent of bY-Z. Setting the random variable W equal to W=bY-Z then by generalised addition law of probability one requires

$$P(X \le W) = \int_w P(X \le w \mid W = w) f(w)dw.$$

Since X is independent of W one requires

$$\begin{aligned} P(X \le W) &= \int_w P(X \le w) f(w)dw \\ &= \int_w \Phi(w) f(w)dw \\ &= E[\Phi(W)]. \end{aligned}$$

Going back to (2.2.24) the event of focus X≤bY-Z may now be expressed in an equivalent form

$$\frac{X+Z}{\varphi Y} \le \frac{b}{\varphi}.$$

This event is equivalent for any φ>0. Now choose $\varphi=\sqrt{(1+\theta^2)}$, since with that choice

$$\frac{X+Z}{\varphi} \sim N\left(\frac{\xi}{\varphi},1\right).$$

Finally one invokes the definitive characterisation of a random variable T(m, δ) with a non-central t distribution having m degrees of freedom and non-centrality parameter $\delta = \xi/\sqrt{1+\theta^2}$ of T(δ,m)=U/Y where U and Y are independent with U~N(δ,1) and Y is defined as $Y = \sqrt{\chi_m^2/m}$. Hence one requires (X+Z)/[Yφ]~T(ξ/φ,m). Finally, for the special case of setting -Z=-ξ=a and $\theta^2 = 0$, and for any value of a and b one requires

$$E\Phi\{(bY-a\} = \text{probt}[b, m, a]. \qquad (2.2.25)$$

Now, since $ms^2/\sigma^2 \sim \chi_m^2$ it follows from this and (2.2.20) that

$$\mathbf{E}(\Psi \mid t_{1-\alpha\ 2,n_A(r+1)-2}) = probt(t_{1-\alpha\ 2,n_A(r+1)-2} + Z_{1-\beta}, m, t_{1-\alpha\ 2,n_A(r+1)-2}). \qquad (2.2.26)$$

The inequality,

$$\text{probt}(Z_{1-\beta}+a,m,a)<1-\beta, \qquad (2.2.27)$$

is valid for any value a > 0 of the non-centrality parameter, for any $0<\beta<0.5$ and all degrees of freedom m ≥1. To prove this result, apply the (monotonically increasing) inverse function *tinv(•, m , a)* of *probt*(•,*m*,*a*), to each side of (2.2.33) so that one can state that (2.2.33) is true if and only if

$$Z_{1-\beta}+a<\text{tinv}(1-\beta,m,a). \qquad (2.2.28)$$

Note, that again the notation *tinv(•,m,a)* agrees with the notation used in SAS.

As the right hand side of (2.2.28) is monotonically decreasing in the degrees of freedom m (for fixed $0<\beta\leq 0.5$ and $a\geq 0$) the result follows since the right side will tend to the left hand side as m tends to ∞.

Hence if n is chosen to satisfy (2.2.17) it follows from (2.2.26), (2.2.27) and (2.2.28) that the expected power is less than its target

$$\mathbf{E}(\Psi) < 1-\beta. \qquad (2.2.29)$$

Hence (2.2.26) suggests that the deficit in the expected power associated with the finite precision of s when the choice of sample size is given by (2.2.17) may be corrected by choosing n_A to satisfy [Julious, 2002a; Julious and Owen, 2006]

$$1-probt\left(\sqrt{\frac{rn_Ad^2}{(r+1)s^2}}, m, t_{1\ \alpha\ 2,n_A(r+1)\ 2}\right) \geq 1-\beta. \qquad (2.2.30)$$

Hence the following result is indicated if

$$n_A \geq \frac{(r+1)s^2\left[tinv\left(1-\beta, \text{m}, t_{1-\alpha\ 2,n_A(r+1)-2}\right)\right]^2}{rd^2}, \qquad (2.2.31)$$

then

$$\mathbf{E}(\Psi) > 1 - \beta . \qquad (2.2.32)$$

To prove this assertion note that the exact form of (2.2.16) is

$$\Psi = \Phi\left(\sqrt{\frac{r n_A d^2}{(r+1)\sigma^2}} - t_{1-\alpha\ 2, n_A(r+1)-2} \right), \qquad (2.2.33)$$

and hence if n satisfies (2.2.31 then

$$\Psi > \Phi\left((s/\sigma) tinv(1-\beta, m, t_{1-\alpha\ 2, n_A(r+1)-2}) - t_{1-\alpha\ 2, n_A(r+1)-2} \right). \qquad (2.2.34)$$

Applying (2.2.21) to (2.2.34) for the expectation conditional on t implies that

$$\mathbf{E}(\Psi \mid t_{1-\alpha\ 2, n_A(r+1)-2}) > probt\{tinv(1-\beta, m, t_{1-\alpha\ 2, n_A(r+1)-2}), m, t_{1-\alpha\ 2, n_A(r+1)-2}\},$$

and the inequality (2.2.32) follows on noting the identity

$$Probt\{tinv(1-\beta, m, t_{1-\alpha\ 2, n_A(r+1)-2}), m, t_{1-\alpha\ 2, n_A(r+1)-2}\} = 1 - \beta .$$

Note that, similar to the standard derivation given earlier in this chapter, the relevant condition (2.2.31) is an inequality (rather than equality) because n may only take an integer value. Thus for given power 1-β, the sample size n would be chosen to be the least integer value satisfying (2.2.31). Note also an approximate solution for n_A is given with $Z_{1-\alpha/2}$ replacing $t_{1-\alpha 2, n_A(r+1)-2}$ in (2.2.31) so that n would be chosen to be the least integer exceed n_A'' where

$$n_A'' = \frac{(r+1)s^2 [tinv(1-\beta, \mathrm{m}, Z_{1-\alpha\ 2})]^2}{rd^2} . \qquad (2.2.35)$$

This equation may be thought of as a version of (2.2.18) which has been adjusted for uncertainty about the unknown true sampling standard deviation σ.

As relationships (2.2.30) and (2.2.31) both have to be solved by iteration for a given power equation (2.2.35) can be used to provide initial values to start the iteration. From simple empirical observation it seems that an expected power of at least 1-β is ensured through adding 1 to the sample size obtained from (2.2.35).

Exact sample size solutions satisfying (2.2.31) for the case r =1 (equal arms) are given in Table 2.3 for 5% significance (α=0.05) and 90% power (β=0.1) and a range of values of degrees of freedom m and a standardised difference defined as, d/s. The last row of the table corresponds to m=∞ and gives the 'traditional' value corresponding to the assumption that σ^2 is known. This row is calculated from (2.2.4) (and (2.2.17)).

Table 2-3. Sample sizes estimated for different standardised differences and degrees of freedom, m, about s^2 from (2.2.30) and from simulations (in brackets). The final line with "infinite" degrees of freedom is from (2.2.4) and the assumption that the -population variance is being used. The type I error is set at a two sided level of 5% and the type II error is set at 10%

	Standardised Difference					
m	0.05	0.10	0.25	0.50	0.75	1.00
10	10933 (10939)	2734 (2751)	439 (439)	111 (111)	50 (50)	29 (29)
25	9327 (9331)	2333 (2335)	374 (374)	95 (95)	43 (43)	25 (25)
50	8853 (8865)	2214 (2217)	355 (355)	90 (90)	41 (41)	23 (23)
75	8702 (8700)	2176 (2182)	349 (349)	88 (88)	40 (40)	23 (23)
100	8627 (8631)	2158 (2159)	346 (347)	88 (88)	40 (40)	23 (23)
200	8516 (8523)	2130 (2132)	342 (342)	86 (86)	39 (39)	23 (23)
500	8451 (8453)	2114 (2114)	339 (339)	86 (86)	39 (39)	22 (22)
∞	8407	2103	337	85	39	22

It is worth noting when considering the approximate formula for n_A of (2.2.35) and (2.2.18) that the ratio of these depends on α, β and m but not on r, s or d. Some values are given in Table 2.4. This table could be used to provide multiplication or inflation factors when standard formulae are used to calculate the sample size, such as (2.3.3), (2.2.4) or (2.2.6), to account for the imprecision in the variance.

$$\text{Inflation Factor (IF)} = \frac{(r+1)s^2\left[tinv(1-\beta, \text{m}, Z_{1-\alpha/2})\right]^2}{\left[Z_{1-\beta} + Z_{1-\alpha/2}\right]^2} \qquad (2.3.36)$$

Note the sample sizes from (2.2.30) converge to (2.2.4) as n_A gets large. Earlier in this chapter, using (2.2.6), the sample sizes were also derived from a non-central t distribution. However, as (2.2.30) converges to (2.2.4) there will be instances for large m where (2.3.40) gives a sample size one smaller than (2.2.6). For the inflation factors though as $Z_{1-\alpha/2}$ not $t_{1-\alpha/2, n_A(r+1)-2}$ is used (2.2.6) becomes (2.2.4). Hence, the inflation factors hold regardless of the original sample size calculation as they are large sample results.

To further confirm the above sample sizes, from (2.2.30), simulations were undertaken in the Interactive Matrix Language (IML) in SAS [1985]. In undertaking the simulations for a given sample size, n, and standardised difference (d/s), a sample variance was first generated with m degrees of freedom and then the power was calculated using expression (2.2.16). For a given sample size and degrees of freedom this was repeated 10,000 times and the average power was calculated across all the simulations. The sample size was then iterated until the average power was over 90% for the given standardised difference. This process was undertaken for different degrees of freedom and for different standardised differences. The simulations confirm the results from (2.2.31) and are given in brackets in Table 2.3.

Table 2-4. Multiplication factors for different levels of two sided significance, type II error and degrees of freedom

		Significance Level (α)			
m	β	0.010	0.025	0.050	0.100
5	0.05	2.232	2.145	2.068	1.980
	0.10	1.819	1.761	1.711	1.652
	0.15	1.614	1.571	1.533	1.489
	0.20	1.482	1.449	1.419	1.385
	0.50	1.122	1.120	1.117	1.114
10	0.05	1.488	1.454	1.425	1.392
	0.10	1.346	1.322	1.301	1.276
	0.15	1.268	1.249	1.233	1.214
	0.20	1.215	1.200	1.187	1.172
	0.50	1.056	1.055	1.054	1.053
25	0.05	1.172	1.160	1.150	1.139
	0.10	1.126	1.117	1.109	1.101
	0.15	1.100	1.092	1.086	1.079
	0.20	1.081	1.075	1.070	1.065
	0.50	1.021	1.021	1.021	1.021
50	0.05	1.083	1.077	1.072	1.067
	0.10	1.061	1.057	1.053	1.049
	0.15	1.049	1.045	1.042	1.039
	0.20	1.040	1.037	1.034	1.032
	0.50	1.010	1.010	1.010	1.010
75	0.05	1.054	1.051	1.047	1.044
	0.10	1.040	1.037	1.035	1.032
	0.15	1.032	1.030	1.028	1.026
	0.20	1.026	1.024	1.023	1.021
	0.50	1.007	1.007	1.007	1.007
100	0.05	1.040	1.038	1.035	1.033
	0.10	1.030	1.028	1.026	1.024
	0.15	1.024	1.022	1.021	1.019
	0.20	1.020	1.018	1.017	1.016
	0.50	1.005	1.005	1.005	1.005

2.3.1.7. Comment

The solution to the problem of accounting for the imprecision of the variance used in sample size calculations provided in this dissertation assumes that the source of variation in the estimate of the variance is pure sampling variation. In practice true variances could differ from trial to trial by more than this suggests and so there is a potential limitation to the solution. However, it should be noted that the proposed solution is an improvement on what is commonly done at the moment, where the problem is ignored altogether. Also, the potential limitations become less important

when one considers that the main issue discussed in this dissertation is the situation where the variance estimates are take from small trials. For such trials pure random variability will be a major component of the overall variability.

The issue of heteroscadicity of trials will be discussed in Chapter 6.

2.3.1.8. Worked Example

Suppose the investigator from the worked examples given earlier wished to account for the imprecision in the sample variance estimate in the design of the trial. Remember the clinical effect of interest is a reduction in blood pressure compared to control of 4mmHg (d) with an observed standard deviation from a pilot study 20mmgHg (s) estimated with 10 degrees of freedom. Thus, the standardised difference equates to $\delta = d/\sigma = 4/20 = 0.20$. For the Type I and Type II errors fixed at 5% and 10% respectively Table 2.4 gives a multiplication factor 1.301 for 10 degrees of freedom. Previously the sample size, assuming the variance in the calculations to be a population variance, was estimated using (2.2.6) at 527 patients in each arm of the trial. To account for the imprecision in the sample variance therefore one needs to increase the sample size estimated earlier by 30% to 745 patients per arm. An inversion of this argument would be to say that by assuming that the standard deviation was a population estimate the sample size could be considered to be underestimated by 30%. This underestimation of the sample size would result in a reduction in the anticipated power by 6% to 84%.

It may seem an unrealistic scenario to undertake a large study where the calculations are based on such few degrees around the variance. However, it is not an unknown occurrence. Chapter 1 gave a worked example where not only was the individual trial sensitive to the assumptions made about the variance used in the sample size calculations but the entire clinical program (which was dependent on the results on the individual trial). Chapter 6 has a discussion on adaptive designs and sample size re-estimation which should be considered for such situations.

2.3.1.9. Bayesian Methods

On an intuitive level if prior information is being used to power a study then it leads one to think of Bayesian methods. In Chapter 6 it will be highlighted how (2.2.36) can be extended to the case where an interim analysis is to be used to re-estimate the variance and hence re-calculate the sample size [Julious, 2004e]. This was an extension of the work of Zucker et al [1999, 2002, 2004] who had previously undertaken the calculation through numerical methods. In Chapter 6 it will be highlighted how these numerical methods give the same result as the calculations using a non-central t-distributions [Julious, 2004e].

Zucker et al [1999, 2004] also highlighted how the sample size re-estimation problem (and analogously the problem in this dissertation) can be thought of in terms of a predictive power calculation with a Jeffrey's prior for the variance. However, this approach also utilises numerical methods to solve – however, these methods do match the results in this chapter [Julious, 2004e].

Bayesian methods will be discussed later in this chapter for the situation where a non-zero mean difference may be anticipate for equivalence, non-inferiority and bioequivalence trials. Also, Bayesian methods will be discussed in detail in subsequent chapters on binary data.

2.3.2. Cross-over Trials

2.3.2.1. Sample Sizes Estimated Assuming the Population Variance to be Known

For the analysis of cross-over trial data this chapter will concentrate on the case where an analysis of variance is the primary analysis, fitting terms for subject, period and treatment. The assumption is that one is undertaking an AB/BA cross-over trial although the methodology described can be extended to a pair wise comparison in a multi-period cross-over trial (with appropriate adjustment to the degrees of freedom). The within subject residual errors are assumed to be sampled from a Normal distribution. This approach is equivalent to the period-adjusted t-test [Senn, 1993].

2.3.2.2. Paired t-tests and Period Adjusted t-tests

The difference between a period adjusted t-test and a standard paired t-test is that for a paired t-test one simply places the observed individual effects on the two treatments in two columns – ignoring any ordering. For each subject a treatment difference is calculated and consequently a mean of these differences, $\bar{d}$, equivalent to $\mu_A - \mu_B$, and a standard deviation of the differences σ_d. The test statistic is thus $\bar{d}\sqrt{n}/\sigma_d$ which is compared to the t distribution on n-1 degrees of freedom.

In a period adjusted t-test for each treatment sequence (AB or BA) a mean difference is calculated, $\bar{d}_{AB}$, equivalent to $\mu_A - \mu_B$, and $\bar{d}_{BA}$, equivalent to $\mu_B - \mu_A$. Assuming that the allocation to each sequence, $n_{AB} = n_{BA} = n/2$, and the within sequence variances, $\sigma^2_{d_{AB}} = \sigma^2_{d_{BA}} = \sigma^2_d$, are equal then the mean difference of interest, $(\bar{d}_{AB} - \bar{d}_{BA})/2$, has the variance $\sigma^2_d(1\ n_{AB} + 1\ n_{AB})/4 = \sigma^2_d/\sqrt{n}$. Thus, the test statistics is

$$\frac{1\ 2\left(\bar{d}_{AB} - \bar{d}_{BA}\right)}{\sigma_d\ \sqrt{n}},$$

which is compared to the t distribution on n-2 degrees of freedom.

If there is truly no period effect,

$$\frac{1/2\left(\bar{d}_{AB}-\bar{d}_{BA}\right)}{\sigma_d/\sqrt{n}} \approx \frac{1/2((\mu_A-\mu_B)-(\mu_B-\mu_A))}{\sigma_d/\sqrt{n}} \approx \frac{\bar{d}\sqrt{n}}{\sigma_d}$$

and thus one would have an equivalent test to a paired t-test but with one less degree of freedom.

2.3.2.3. Sample Size Calculations

To estimate a sample size for a cross-over trials as well as quantifying the within subject estimate of the difference in treatment means that is of interest, the effect size, one also needs an estimate of the within- (intra-) subject standard deviation σ_w. The within-subject standard deviation is taken from the residual line of an ANOVA model and quantifies the expected variation among repeated measurements on the same individual [Julious, Campbell and Altman, 1999]. With an estimate of both the within subject standard deviation and the effect size (2.2.1) can again be solved as per a parallel group study

$$n = \frac{2\left(Z_{1-\beta}+Z_{1-\alpha/2}\right)^2\sigma_w^2}{d^2}, \qquad (2.2.36)$$

where n here is the total sample size. Note that the allocation ratio has not been used as per (2.2.3) as in a cross-over trial the meaning of r would be the allocation ratio per treatment sequence AB and BA. The assumption here is that subjects will be equally assigned to each sequence. If a sample variance is to be used in the analysis then one can rewrite (2.2.36) as

$$n \geq \frac{2\left(Z_{1-\beta}+t_{1-\alpha/2,n-2}\right)^2\sigma_w^2}{d^2}, \qquad (2.2.37)$$

which in turn can be rewritten in terms of power to solve iteratively for n

$$1-\beta = \Phi\left(\sqrt{\frac{nd^2}{2\sigma_w^2}} - t_{1-\alpha/2,n-2}\right). \qquad (2.2.38).$$

Similarly to parallel group trials, when the population variances is considered unknown for the statistical analysis, under H_1: $d \neq 0$ the Type II error (and hence the power) should be calculated under the assumption of a non-central t distribution with degrees of freedom n-2 and non-centrality parameter $\sqrt{nd^2/2\sigma_w^2}$ [Senn, 1993; Kupper and Hafner, 1989, Julious, 2004a]. Thus, (2.2.44) can be rewritten as

$$1-\beta = 1-\text{probt}\left(t_{1-\alpha/2,n-2},\ n-2,\ \sqrt{\frac{nd^2}{2\sigma_w^2}}\right). \qquad (2.2.39)$$

In the same manner to a parallel group study one can add a correction factor of $Z_{1-\alpha/2}/2$ to (2.2.36) to allow for the Normal approximation, and use this for initial calculations in (2.2.39) [Guenther, 1981]

$$n = \frac{2\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 \sigma_w^2}{d^2} + \frac{Z_{1-\alpha/2}^2}{2}. \qquad (2.2.40)$$

For quick calculations one can adapt (2.2.36) for the calculation of sample sizes with 90% power and a two-sided 5% type I error rate

$$n = \frac{21\sigma_w^2}{d^2}. \qquad (2.2.41)$$

As with the parallel group equivalents (2.2.40) and (2.2.41) give slightly lower results than (2.2.39), while also compared to (2.2.39), (2.2.38) will mostly give the same sample size – occasionally underestimating by 1. Table 2.5 gives sample sizes using (2.2.39) for various standardised differences ($\delta = d/\sigma$).

Table 2-5. Total sample sizes for a cross-over study for different standardised differences for 90% power and two sided type I error rate of 5%.

δ	n
0.10	2104
0.20	528
0.30	236
0.40	134
0.50	87
0.60	61
0.65	52
0.70	45
0.80	35
0.90	29
1.00	24

The total sample sizes for cross-over trials are nearly the equivalent to that for one arm of parallel group studies, for each standardised difference (δ). The slight differences are accounted for by the different degrees of freedom used in (2.2.6) and (2.2.39). Practically, though, they are the same. It should be noted, however, that the standardised differences in Tables 2.1 and 2.5 represent different quantities. The within- subject variance in a cross-over trial can be derived from $\sigma_w^2 = \sigma^2(1-\rho)$, where σ^2 is the population variance from a conventional parallel group design and ρ is the Pearson correlation coefficient estimated between two measures on the same subject. For a relatively modest correlation of 0.5, the within-subject variance would be half the population variance, and as a consequence the equivalent standardised difference would be 40% larger in a cross-over compared to a parallel group study. Parallel group and

cross-over trials will only have an equivalent standardised difference for a zero correlation.

2.3.2.4. Worked Example

An investigator wishes to design a hypertension trial similar to that in the section on parallel group trials where the clinical effect of interest is a reduction in blood pressure compared to control of 4mmHg (d). The expected within-subject standard deviation in the trial population the trial is expected to be half that of the between-subject standard deviation at 8mmgHg (σ). Thus, the standardised difference equates to $\delta = d/\sigma = 4/8 = 0.50$. For the Type I and Type II errors fixed at 5% and 10% respectively Table 2.5 gives a total sample size of 87.

2.3.2.5. Sensitivity Analysis About the Variance Used in the Sample Size Calculations

As with parallel group trials the sensitivity of the sample size estimate in a cross-over trial is relatively straightforward to investigate. Equation (2.2.9) can be used with an estimate of the within subject population variance.

Table 2.6 gives the total sample size required for different standardised differences for a cross-over superiority trial along with the sensitivity of these sample sizes to the 95th percentile of the variance for different degrees of freedom. From Table 2.6 it seems, like with the parallel group trials, that to ensure that 50% and 80% power with the 95th percentile about the variance one would require 10 and 75 degrees of freedom respectively.

2.3.2.6. Worked Example

Following on from the worked example given earlier – a hypertension trial similar to that where the clinical effect of interest is a reduction in blood pressure of 10mmHg (d). Suppose that the within-subject standard deviation, of 20mmgHg (σ), was estimated with 20 degrees of freedom then from Table 2.9 with an estimated sample size of 87 for the trial one would have 67% power if the actual variance was nearer the plausibly high value.

Table 2-6. Total sample sizes for a superiority cross-over trial for different standardised differences with 90% power and 5% type I error rate along with the power corresponding to the 95th percentile of the variance for different degrees of freedom

		Degrees of Freedom										
δ	n	5	10	15	20	25	30	40	50	75	100	200
0.10	2104	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.84
0.20	528	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.85
0.30	236	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.85
0.40	134	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.85
0.50	87	0.34	0.53	0.62	0.67	0.70	0.73	0.76	0.78	0.80	0.82	0.85
0.60	61	0.34	0.53	0.62	0.67	0.70	0.73	0.76	0.78	0.80	0.82	0.85
0.70	45	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.85
0.80	35	0.34	0.53	0.62	0.67	0.70	0.72	0.75	0.77	0.80	0.82	0.85
0.90	29	0.35	0.55	0.63	0.68	0.71	0.74	0.77	0.79	0.81	0.83	0.86
1.00	24	0.35	0.55	0.63	0.68	0.72	0.74	0.77	0.79	0.82	0.83	0.86

2.3.2.7. Calculations taking Account of the Imprecision of the Variance Used in the Sample Size Calculations

To account for the imprecision of the variance used in the sample size calculations the results for parallel group trials can be generalised to give the following formula

$$n \geq \frac{2s_w^2\left[tinv\left(1-\beta, \mathrm{m}, t_{1-\alpha\ 2,n-2}\right)\right]^2}{d^2}. \qquad (2.2.42)$$

Where n is the least integer value for (2.2.42) to hold. One can rewrite (2.2.42) in terms of power to obtain the following result

$$1-\beta = 1-probt\left(\sqrt{\frac{nd^2}{2s_w^2}}, m, t_{1-\alpha\ 2,n-2}\right) \qquad (2.2.43)$$

Replacing the t-statistic with a Z-statistic gives one the following result

$$n = \frac{2s_w^2\left[tinv\left(1-\beta, \mathrm{m}, \mathrm{Z}_{1-\alpha\ 2}\right)\right]^2}{d^2} \qquad (2.2.44)$$

which allows one to have a direct estimate of the sample size and also gives an initial value for iterations for (2.2.42).

Exact sample size solutions satisfying (2.2.42) are given in Table 2.7 for 5% significance (α=0.05) and 90% power (β=0.1) and a range of values of degrees of freedom m and a standardised difference defined as, d/s. The last row of the table corresponds to m = ∞ and gives the 'traditional' value corresponding to the assumption that σ^2 is known from (2.2.38).

Table 2-7. Sample sizes estimated for different standardised differences and degrees of freedom from (2.2.42). The final line with "infinite" degrees of freedom is from (2.2.38) and the assumption that the population variance is being used. The type I error is set at a two sided significance level of 5% and the type II error is set at 10%

	Standardised Difference					
m	0.05	0.10	0.25	0.50	0.75	1.00
10	10934	2735	439	111	51	30
25	9328	2333	375	95	44	25
50	8854	2215	356	91	41	24
75	8703	2177	350	89	41	24
100	8628	2158	347	88	40	24
200	8517	2131	343	87	40	23
500	8451	2114	340	86	40	23
∞	8408	2103	338	86	39	23

As with parallel group trials multiplication factors derived from (2.2.44) and (2.2.36) can be obtained to assist in trial design. As these depend only α, β and m (but not on r, s or d) these ratios are the same as, and are given in, Table 2.4.

2.4. Equivalence Trials

2.4.1. Parallel Group Trials

2.4.1.1. Sample Sizes Estimated Assuming the Population Variance to be Known

2.4.1.2. General Case

Recall from Chapter 1 that the total Type II error (define as $\beta = \beta_1 + \beta_2$) is derived from the following result

$$Z_{1-\beta_1} = \frac{-d-\Delta}{\sqrt{Var(S)}} - Z_{1-\alpha} \text{ and } Z_{1-\beta_2} = \frac{d-\Delta}{\sqrt{Var(S)}} - Z_{1-\alpha}. \qquad (2.3.1)$$

For equivalence trials for the general case where the expected true mean difference is not fixed to be zero the sample size cannot be derived directly. This is because the total Type II error is the sum of the Type II errors associated with each one-tailed test. As is the case with superiority trials Var(S) can be defined as

$$Var(S) = \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B} = \frac{r+1}{r} \cdot \frac{\sigma^2}{n_A}. \qquad (2.3.2)$$

From this, and the fact that $\beta = \beta_1 + \beta_2$, the following can be used to derive the Type II error (and power)

$$1-\beta = \Phi\left(\sqrt{\frac{((\mu_A-\mu_B)-d)^2 rn_A}{(r+1)\sigma^2}} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{\frac{((\mu_A-\mu_B)+d)^2 rn_B}{(r+1)\sigma^2}} - Z_{1-\alpha}\right) - 1.$$

The sample size cannot be derived directly; instead one has to iterate until a sample size is reached which gives the required Type II error (and power). If the variance is to be considered unknown for the statistical analysis (2.3.4) can be used

$$1-\beta=\Phi\left(\sqrt{\frac{((\mu_A-\mu_B)-d)^2 rn_A}{(r+1)\sigma^2}}-t_{1-\alpha,n_A(r+1)-2}\right)+\Phi\left(\sqrt{\frac{((\mu_A-\mu_B)+d)^2 rn_A}{(r+1)\sigma^2}}-t_{1-\alpha,n_A(r+1)-2}\right)-1. \qquad (2.3.3)$$

As with superiority trials it is best to assume a non-central t-distribution to calculate the Type II error and power. Under the assumption of a non-central t-distribution the power can be calculated using the following [Owen, 1965; Diletti, Hauschke, Steinijans, 1991; Julious, 2004a]

$$1-\beta=\text{probt}\left(-t_{1-\alpha,n_A(r+1)-2},n_A(r+1)-2,\tau_2\right)-\text{probt}\left(t_{1-\alpha,n_A(r+1)-2},n_A(r+1)-2,\tau_1\right), \qquad (2.3.4)$$

where τ_1 and τ_2 are non centrality parameters defined as

$$\tau_1=\frac{((\mu_A-\mu_B)+d)\sqrt{rn_A}}{\sqrt{(r+1)\sigma^2}} \text{ and } \tau_2=\frac{((\mu_A-\mu_B)-d)\sqrt{rn_A}}{\sqrt{(r+1)\sigma^2}}.$$

For quick calculations, and to provide an initial value for the sample size in the iterations, an estimate of the sample size can be obtained from the following

$$n_A=\frac{(r+1)\sigma^2\left(Z_{1-\beta}+Z_{1-\alpha}\right)^2}{r((\mu_A-\mu_B)-d)^2}. \qquad (2.3.5)$$

This provides reasonable approximations for case of $\mu_A-\mu_B>0$, especially when the mean difference approaches d. For very quick calculations, for 90% power and Type I error of 2.5%, the following formula can be used

$$n_A=\frac{10.5\sigma^2(r+1)}{((\mu_A-\mu_B)-d)^2 r}, \qquad (2.3.6)$$

or for r=1

$$n_A=\frac{21\sigma^2}{((\mu_A-\mu_B)-d)^2}. \qquad (2.3.7)$$

2.4.1.3. Special Case of No Treatment Difference

For the special case of no treatment difference, $\mu_A-\mu_B=0$, (2.3.3) can be rewritten to obtain a direct estimate of the sample size

$$n_A=\frac{(r+1)\sigma^2\left(Z_{1-\beta/2}+Z_{1-\alpha}\right)^2}{rd^2}. \qquad (2.3.8)$$

For case of the variance considered unknown for the statistical analysis, (2.3.8) can be written in terms of

$$n_A = \frac{(r+1)\sigma^2\left(Z_{1-\beta/2} + t_{1-\alpha,n_A(r+1)-2}\right)^2}{rd^2}. \qquad (2.3.9)$$

Equation (2.3.9) can be rewritten to give power in terms of the sample size

$$1-\beta = 2\Phi\left(\sqrt{\frac{rd^2 n_1}{(r+1)\sigma^2}} - t_{1-\alpha,n_A(r+1)-2}\right) - 1, \qquad (2.3.10)$$

and, similarly to (2.3.4), under the assumption of a non-central t-distribution, the power can be derived from

$$1-\beta = 2\text{probt}\left(-t_{1-\alpha,n_A(r+1)-2}, n_A(r+1)-2, \tau\right) - 1, \qquad (2.3.11)$$

where τ is defined as

$$\tau = \frac{-\sqrt{n_A}rd}{\sqrt{(r+1)\sigma^2}}.$$

For quick calculations, for 90% power and Type I error of 2.5%, the following formula, similar to (2.3.6) can be used

$$n_A = \frac{13\sigma^2(r+1)}{d^2 r}, \qquad (2.3.12)$$

or, for r=1,

$$n_A = \frac{26\sigma^2}{d^2}. \qquad (2.3.13)$$

The quick equations give reasonable estimates of the sample size, underestimating by one or two, and thus provide reasonable initial values for (2.3.4) and (2.3.11). It was worth noticing the difference between (2.3.12) and (2.3.13) compared to (2.3.6) and (2.3.7). The difference in the coefficients (10.5 and 21 compared to 13 and 26) is to do with the non-symmetric allocation of the Type II error if the population mean is non-zero. Table 2.8 gives sample sizes for equivalence trials using (2.3.4).

Table 2-8. Sample sizes (n_A) for one arm of a parallel group equivalence study with equal allocation (r=1) for different standardised equivalence limits and true mean differences (as a percentage of the equivalence limit) for 90% power and type I error rate of 2.5%

	Percentage Mean Difference				
d	0%	10%	15%	20%	25%
0.10	2600	2762	2980	3306	3741
0.20	651	691	746	827	936
0.30	290	308	332	369	417
0.40	164	174	188	208	235
0.50	105	112	121	134	151
0.60	74	78	84	93	105
0.70	55	58	62	69	78
0.80	42	45	48	53	60
0.90	34	36	38	42	48
1.00	27	29	31	35	39

2.4.1.4. Worked Example

An investigator wishes to design a hypertension trial where the objective is to demonstrate equivalence between the two treatments. The largest clinically acceptable effect for which equivalence can be declared is a change in blood pressure of 4mmHg (d). There is to be equal allocation between groups. The true mean difference between the treatments is thought to be zero and the expected standard deviation in the population in which the trial is to be undertaken is 25mmHg (σ). Thus, the standardised equivalence limits equate to $\pm\delta = \pm d/\sigma = \pm 5/25 = \pm 0.20$. For the Type I and Type II errors fixed at 2.5% and 10% respectively Table 2.8 gives a sample size of 651 patients in each arm of the trial.

Suppose the true mean difference is thought to be 1mmHg. This equates to 20% of the standardised equivalence limits and would inflate the sample size to 827 patients in each arm of the trial.

2.4.1.5. Sensitivity Analysis About the Variance Used in the Sample Size Calculations

As with superiority trials described earlier in this chapter the sensitivity of the sample size estimate to the variance used in the calculations is relatively straightforward to investigate. Equation (2.2.9) can be used to estimate a plausibly large value for the population variance and from this the sensitivity of the study (assessed as a loss in power) to this high variance.

With equivalence trials however, one has the further factor to investigate of the sensitivity of calculations about the true mean difference. If one has assumed no difference when it was truly non-zero then this will have an effect on the power of the study.

Table 2.9 gives the sample size per group required for different standardised equivalence limits for a parallel group equivalence trial along with the sensitivity of

these sample sizes to the 95th percentile of the variance (for different degrees of freedom) and different mean differences (assuming the mean difference is zero).

Table 2-9. Sample sizes per arm for a parallel group equivalence trial for different standardised equivalence limits with 90% power and 2.5% type I error rate along with the power corresponding to the 95th percentile of the variance for different degrees of freedom and different mean differences

d	Sample Size	True Mean Diff (%)	Degrees of Freedom 10	25	50	75	100	500	∞
0.05	10397	0	0.24	0.57	0.70	0.75	0.78	0.85	0.90
		5	0.24	0.57	0.70	0.75	0.77	0.85	0.89
		10	0.23	0.56	0.68	0.73	0.76	0.83	0.88
		15	0.22	0.54	0.66	0.70	0.73	0.80	0.85
		20	0.22	0.51	0.62	0.67	0.69	0.76	0.81
		25	0.20	0.47	0.58	0.62	0.64	0.72	0.77
0.10	2600	0	0.24	0.57	0.70	0.75	0.78	0.85	0.90
		5	0.24	0.57	0.70	0.75	0.77	0.85	0.89
		10	0.23	0.56	0.68	0.73	0.76	0.83	0.88
		15	0.22	0.54	0.66	0.70	0.73	0.80	0.85
		20	0.22	0.51	0.62	0.67	0.69	0.76	0.81
		25	0.20	0.47	0.58	0.62	0.64	0.72	0.77
0.15	1157	0	0.24	0.57	0.70	0.75	0.78	0.85	0.90
		5	0.24	0.57	0.70	0.75	0.77	0.85	0.89
		10	0.23	0.56	0.68	0.73	0.76	0.83	0.88
		15	0.22	0.54	0.66	0.70	0.73	0.80	0.85
		20	0.22	0.51	0.62	0.67	0.69	0.76	0.81
		25	0.20	0.47	0.58	0.62	0.64	0.72	0.77
0.20	651	0	0.24	0.57	0.70	0.75	0.78	0.85	0.90
		5	0.24	0.57	0.70	0.75	0.77	0.85	0.89
		10	0.23	0.56	0.68	0.73	0.75	0.83	0.88
		15	0.22	0.54	0.66	0.70	0.73	0.80	0.85
		20	0.21	0.51	0.62	0.66	0.69	0.76	0.81
		25	0.20	0.47	0.58	0.62	0.64	0.72	0.77
0.25	417	0	0.24	0.57	0.70	0.75	0.78	0.85	0.90
		5	0.23	0.57	0.70	0.75	0.77	0.85	0.89
		10	0.23	0.56	0.68	0.73	0.75	0.83	0.88
		15	0.22	0.53	0.66	0.70	0.73	0.80	0.85
		20	0.21	0.51	0.62	0.66	0.69	0.76	0.81
		25	0.20	0.47	0.58	0.62	0.64	0.72	0.77

From Table 2.9 it seems that assuming a trial has been designed with 90% power:

- With 25 degrees of freedom, if the true variance was nearer to the 95th percentile then one would have 50% power. With 100 degrees of freedom then if the true variance was nearer to the 95th percentile then one would have 80% power. Both calculations assume there is no mean difference.

- If one's variance estimate was close to the true variance then even if the true mean difference was 20% of the standardised equivalence limit then one would still have 80% in the study.

- With moderate degrees of freedom a study is reasonably robust to deviations in the assumptions either about the true mean or the variance but less so to deviations in both simultaneously.

2.4.1.6. Worked Example

In the worked example given earlier a sample size of 651 was calculated for a standardised equivalence limit of 0.20. Now suppose the variance used in the calculations was estimate with 25 degrees of freedom then from Table 2.9 Table 2.10 could be built as an investigation of the sensitivity of the study.

Table 2-10. Worked example of a sensitivity analysis for an individual equivalence study.

True Difference (%)	Power
0	0.57
5	0.57
10	0.56
15	0.54
20	0.51
25	0.47

The values in each cell are the calculated powers of the study for the different scenarios.

2.4.1.7. Calculations Taking Account of the Imprecision of the Variances Used in the Sample Size Calculations

2.4.1.8. General Case

Extending the arguments from superiority trials. To account for the degrees of freedom of the sample variance used in the calculations the following equation could be used to calculate the power

$$1-\beta = \text{probt}\left(-\tau_2, m, -t_{1-\alpha, n_A(r+1)-2}\right) - \text{probt}\left(\tau_1, m, t_{1-\alpha, n_A(r+1)-2}\right), \quad (2.3.14)$$

where τ_1 and τ_2 are the absolute standardised equivalence limits defined as defined as

$$\tau_1 = \frac{|(\mu_A - \mu_B) - d|\sqrt{rn_A}}{\sqrt{(r+1)s^2}} \text{ and } \tau_2 = \frac{|(\mu_A - \mu_B) + d|\sqrt{rn_A}}{\sqrt{(r+1)s^2}}.$$

To calculate the sample size one would need to iterate to find the minimum value that would give the required power from (2.3.14).

For non zero treatment differences (i.e. for $\mu_A - \mu_B > 0$) most of the Type Two error would be coming from just one tail and hence the power could be estimated from

$$1-\beta = 1 - \text{probt}\left(\frac{|(\mu_A - \mu_B) - d|\sqrt{rn_A}}{\sqrt{(r+1)s^2}}, m, t_{1-\alpha, n_A(r+1)-2} \right). \qquad (2.3.15)$$

Which when written in terms of n becomes

$$n_A \geq \frac{(r+1)s^2\left[tinv(1-\beta, \text{m}, t_{1-\alpha, n_A(r+1)-2})\right]^2}{r((\mu_A - \mu_B) - d)^2}. \qquad (2.3.16)$$

Replacing the t-statistic with a Z-statistic and (2.3.15) can in turn be approximated from the following equation to give a direct estimate of the sample size

$$n_A = \frac{(r+1)s^2\left[tinv(1-\beta, \text{m}, Z_{1-\alpha})\right]^2}{r((\mu_A - \mu_B) - d)^2}. \qquad (2.3.17)$$

This direct estimate could be used to provide initial estimates of the sample size for (2.3.14).

2.4.1.9. Special Case of No Treatment Difference

For the special case of no treatment difference the power can be estimated from

$$1-\beta = 2\text{probt}(\tau, m, -t_{1-\alpha, n_A(r+1)-2}) - 1, \qquad (2.3.18)$$

where τ is defined as

$$\tau = \frac{-\sqrt{n_A}rd}{\sqrt{(r+1)s^2}},$$

which when written in terms of n becomes

$$n_A \geq \frac{(r+1)s^2\left[tinv(1-\beta\ 2, \text{m}, t_{1-\alpha, n_A(r+1)-2})\right]^2}{rd^2}. \qquad (2.3.19)$$

Replacing the t-statistic with a Z-statistic and (2.3.19) can in turn be approximated from the following equation to give a direct estimate of the sample size.

$$n_A = \frac{(r+1)s^2\left[tinv(1-\beta\ 2, \text{m}, Z_{1-\alpha})\right]^2}{rd^2} \qquad (2.3.20)$$

Tables 2.11 and 2.12 are produced for the special case of no mean difference between treatments. Table 2.11 gives the sample sizes for different degrees of freedom and standardised equivalence limits.

Table 2-11. Sample sizes estimated for different standardised equivalence limits and degrees of freedom from (2.3.14). The final column with "infinite" degrees of freedom is from (2.3.3) and the assumption that the population variance is being used. The type I error is set at a two one-sided significance level of 2.5% and the type II error is set at 10%

	Degrees of Freedom						
d	10	25	50	75	100	500	∞
0.10	3705	2990	2787	2723	2692	2618	2600
0.20	927	748	698	682	674	655	651
0.30	413	333	311	304	300	292	290
0.40	233	188	175	171	169	165	164
0.50	149	121	113	110	109	106	105
0.60	104	84	79	77	76	74	73
0.70	77	62	58	57	56	55	54
0.80	59	48	45	44	43	42	42
0.90	47	38	36	35	34	33	33
1.00	38	31	29	28	28	27	27

Table 2.12 gives the multiplication factors, compared to assuming one has the population variance, for various degrees of freedom and Type I and II errors. Similar to superiority trials (2.3.20) converges to (2.3.10), however, the multiplication factors can be used regardless of the original formula for calculations.

2.4.1.10. Worked Example

Returning to the worked example given earlier for a standardised equivalence limit of 0.20 with 25 degrees of freedom for the variance Table 2.11 gives a sample size of 748. This compares to sample size of 651 calculated assuming one had the population variance for calculations - a potential under estimation of the sample size of 16%.

Table 2-12. Multiplication factors for different levels of one sided significance, type II error and degrees of freedom

m	β	Significance Level (α) 0.010	0.025	0.050	0.100
5	0.05	2.649	2.509	2.385	2.238
	0.10	2.167	2.068	1.980	1.875
	0.15	1.929	1.850	1.780	1.696
	0.20	1.776	1.711	1.652	1.581
	0.50	1.367	1.337	1.311	1.278
10	0.05	1.611	1.562	1.520	1.470
	0.10	1.463	1.425	1.392	1.353
	0.15	1.382	1.351	1.323	1.290
	0.20	1.328	1.301	1.276	1.248
	0.50	1.166	1.153	1.141	1.127
25	0.05	1.208	1.192	1.178	1.162
	0.10	1.163	1.150	1.139	1.125
	0.15	1.137	1.126	1.116	1.105
	0.20	1.119	1.109	1.101	1.091
	0.50	1.062	1.058	1.053	1.058
50	0.05	1.099	1.091	1.085	1.077
	0.10	1.078	1.072	1.067	1.060
	0.15	1.066	1.061	1.056	1.051
	0.20	1.058	1.053	1.049	1.044
	0.50	1.031	1.028	1.026	1.024
75	0.05	1.065	1.060	1.056	1.051
	0.10	1.052	1.047	1.044	1.040
	0.15	1.044	1.040	1.037	1.033
	0.20	1.038	1.035	1.032	1.029
	0.50	1.020	1.019	1.017	1.016
100	0.05	1.048	1.044	1.041	1.038
	0.10	1.038	1.035	1.033	1.030
	0.15	1.033	1.030	1.028	1.025
	0.20	1.029	1.026	1.024	1.022
	0.50	1.015	1.014	1.013	1.012

2.4.1.11. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations

As well as being sensitive to the assumptions around the variance the sample size calculations are also sensitive to the assumptions around the assumed mean difference. The more the mean difference deviates away from the assumptions in the calculations (and nearer the margins) the greater the reduction in power. If one has an estimate of the mean difference ($\bar{x}_A - \bar{x}_B$) from a previous study then one could use the standard error around this mean difference ($se(\bar{x}_A - \bar{x}_B)$) when estimating the sample size.

For the general case $\bar{x}_A - \bar{x}_B \neq 0$ the power for a given sample size can be estimated (through numerical integration) from

$$1-\beta=\sum_{p=0.001}^{0.998}0.001x\left(\left[\begin{array}{l}\text{probt}\left(\dfrac{\sqrt{rn_A}\left((\bar{x}_A-\bar{x}_B)+Z_p se(\bar{x}_A-\bar{x}_B)-d\right)}{\sqrt{(r+1)s^2}},m,-t_{1-\alpha,n_A(r+1)-2}\right)\\ -\text{probt}\left(\dfrac{\sqrt{rn_A}\left((\bar{x}_A-\bar{x}_B)+Z_p se(\bar{x}_A-\bar{x}_B)+d\right)}{\sqrt{(r+1)s^2}},m,t_{1-\alpha,n_A(r+1)-2}\right)\end{array}\right]+\left[\begin{array}{l}\text{probt}\left(\dfrac{\sqrt{rn_A}\left((\bar{x}_A-\bar{x}_B)+Z_{p+0.001} se(\bar{x}_A-\bar{x}_B)-d\right)}{\sqrt{(r+1)s^2}},m,-t_{1-\alpha,n_A(r+1)-2}\right)\\ -\text{probt}\left(\dfrac{\sqrt{rn_A}\left((\bar{x}_A-\bar{x}_B)+Z_{p+0.001} se(\bar{x}_A-\bar{x}_B)+d\right)}{\sqrt{(r+1)s^2}},m,t_{1-\alpha,n_A(r+1)-2}\right)\end{array}\right]\right)/2, \qquad (2.3.21)$$

where the sample size required is the minimum value which gives the required sample size. Z_p is the value from the Normal distribution that equates to the percentile p.

This result however is practically unappealing as the only way for a sample size to be easily estimable is for the cases where equivalence has already been demonstrated i.e. all plausible values for the mean difference fall within the equivalence margin. Practically to allow for possible mean differences between treatments one should first investigate the sensitivity of the trial to the assumptions about the mean difference (all trials will still have 80% power if one assumes no mean difference for the calculations but the true mean difference is 20% of the equivalence margin). If the trial is quite sensitive to the mean difference assumption then the calculations should maybe be repeated allowing for a small mean difference between treatments.

2.4.1.12. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations – A Bayesian Approach

In this section it will be demonstrated how simple Bayesian methods could be employed to estimate sample sizes allowing for imprecision in the mean and variance. The rationale for this section is not to give a detailed exposition of Bayesian methodology but to detail how this methodology may be applied.

The context of the methodology with the situation in hand is to interrogate sample sizes where an equivalence study is to be planned and a mean difference, $d_1=\bar{x}_A-\bar{x}_B$, has previously been observed. In the prospective trial being designed inference is to be made about the 'true' difference θ.

For the given sample size what needs to be determined is the probability of observing a given difference, d_1, or greater for θ given that d_1 has already been observed i.e. $\text{Prob}(\theta > d_1|d_1)$.

Note, the methodology described in the previous subsection, (2.3.1.11), could be considered to be sample calculations calculated under a Bayesian framework but with a non-informative prior distribution for θ.

What is to be undertaken now is an investigation of calculations where the assumption is that there is prior knowledge about the mean difference that will be used. Simply, for

a hypothesis H suppose that prior to observing data D one believed H with probability P(H) then after observing D one should believe H with probability (from Bayes' rule)

$$P(H|D)=P(D|H)P(H)/P(D). \quad (2.3.22)$$

For inference about an unknown continuous parameter, θ, what one is interested in is how would our belief about θ change. If the prior is expressed in the density $p(\theta)$ and if subsequently data x are observed then the posterior distribution is expressed in the density, $p(\theta|x)$, where the Bayes' rule for densities is

$$p(\theta|x)\propto\lambda(x|\theta)p(\theta), \quad (2.3.23)$$

where $\lambda(x|\theta)$ is the likelihood function. Note the constant of proportionality is chosen so that it integrates to 1 and the likelihood function is considered defined for the whole range (though may be zero for parts of it).

For Normal data of the form

$$X_1,X_2,...X_n\sim N(\theta,\sigma^2), \quad (2.3.24)$$

where one wishes to obtain inference about θ for given σ^2 one has the following prior

$$\theta\sim N(\mu_0,\sigma_0^2), \quad (2.3.25)$$

The Normal family is conjugate in this case (both the prior and the posterior have the same distribution). The Bayesian updating rules for the case described in this section can be defined as follows.

2.4.1.13. Prior Response

Prior values for the mean difference and population standard deviation are defined as d_0 (i.e. for $\bar{x}_A-\bar{x}_B$) and s_0 respectively. These values can be subjective values taken as beliefs about the mean difference.

2.4.1.14. Anticipated Response

The anticipated mean difference and population standard deviation are defined as d_1 and s_1 respectively. These values are taken as objective values observed in a previous clinical trial. Hence $s_1\sqrt{(r+1)/rn}$ is an estimate of the standard deviation around the mean where r is the allocation ratio between groups.

2.4.1.15. Posterior Response

With the anticipated and prior responses the posterior distribution can be calculated through a weighted sum of the prior and anticipated responses. The posterior estimate of the variance around the mean, s_n^2, is defined as

$$s_n^2 = \left(\frac{rn}{s_1^2 (r+1)} + \frac{1}{s_0^2} \right)^{-1} \tag{2.3.26}$$

and the posterior estimate of the mean difference, d_n, is defined as

$$d_n = s_n^2 \left(\frac{d_0}{s_0^2} + \frac{d_1 rn}{s_1^2 (r+1)} \right). \tag{2.3.27}$$

From these posterior values a density distribution for $prob(\theta < d_i \mid d_1)$ can be defined such that a probability of observing d_i, or greater, for a given d_n would be

$$prob(\theta > d_i | d_1) = \Phi\left(\frac{d_n - d_i}{s_n} \right), \tag{2.3.28}$$

or alternatively for $prob(\theta < d_i \mid d_1)$

$$prob(\theta > d_i | d_1) = \Phi\left(\frac{d_i - d_n}{s_n} \right).$$

From this result a mean difference d_i for given percentile p can be estimated and put into the following result for the general case $\bar{x}_A - \bar{x}_B \neq 0$, to estimate the power for a given sample size.

$$1-\beta = \sum_{p=0.0001}^{0.9998} 0.0001 \times \left(\left[\text{probt}\left(\frac{\sqrt{rn}_A \left(-d - d_p\right)}{\sqrt{(r+1)s^2}}, m, -t_{1-\alpha_A (r+1)-2} \right) - \text{probt}\left(\frac{\sqrt{rn}_A \left(d - d_p\right)}{\sqrt{(r+1)s^2}}, m, t_{1-\alpha_A (r+1)-2} \right) \right] + \left[\text{probt}\left(\frac{\sqrt{rn}_A \left(-d - d_{p+0.0001}\right)}{\sqrt{(r+1)s^2}}, m, -t_{1-\alpha_A (r+1)-2} \right) - \text{probt}\left(\frac{\sqrt{rn}_A \left(d - d_{p+0.0001}\right)}{\sqrt{(r+1)s^2}}, m, t_{1-\alpha_A (r+1)-2} \right) \right] \right) / 2, \tag{2.3.29}$$

where the sample size is estimated through numerical integration and iteration.

Note that for this Bayesian approach if s_0 is set to a very large value (2.3.29) will approach (2.3.21) described earlier as

$$s_n^2 = \left(\frac{rn}{s_1^2 (r+1)} + \frac{1}{s_0^2} \right)^{-1} \rightarrow \left(\frac{(r+1)s_1^2}{rn} \right) \text{ and } d_n = s_n^2 \left(\frac{d_0}{s_0^2} + \frac{d_1 rn}{s_1^2 (r+1)} \right) \rightarrow d_1.$$

In the context of the problem here elementary Bayesian procedures can enhance insights into the reliability of inferences. As well as using empirical observation, beliefs about the results that are anticipated can be used in sample size calculations. The invocation of a prior could provide an opportunity to the explore robustness (sensitivity) of calculations. For example you may be more sceptical in your priors than results previously observed. This more sceptical subjective prior could be used to calculate posterior probabilities and hence sample size calculations.

As with the result discussed in section (2.3.1.11) there is no great practical application of (2.3.39). Bayesian inference around the mean response will be discussed in greater detail in Chapter 4 for binary data.

2.4.2. Cross-over Trials

The methodologies and assumptions for an equivalence trial with a cross-over design are the same as those for parallel group equivalence trials (for the methodologies) and a superiority cross-over trials (for assumptions about the parameters). This subsection will therefore only go briefly through the sample size calculations for an equivalence trial with a cross-over design.

2.4.2.1. Sample Size Estimated Assuming the Population Variance to be Known

2.4.2.2. General Case

The Type II error (and power) can be estimated from

$$1-\beta=\Phi\left(\sqrt{\frac{((\mu_A-\mu_B)-d)^2 n}{2\sigma_w^2}}-Z_{1-\alpha}\right)+\Phi\left(\sqrt{\frac{((\mu_A-\mu_B)+d)^2 n}{2\sigma_w^2}}-Z_{1-\alpha}\right)-1. \qquad (2.3.30)$$

If the variance is to be considered unknown for the statistical analysis then (2.3.22) can be rewritten as

$$1-\beta=\Phi\left(\sqrt{\frac{((\mu_A-\mu_B)-d)^2 n}{2\sigma_w^2}}-t_{1-\alpha,n-2}\right)+\Phi\left(\sqrt{\frac{((\mu_A-\mu_B)+d)^2 n}{2\sigma_w^2}}-t_{1-\alpha,n-2}\right)-1, \qquad (2.3.31)$$

and under the assumption of a non-central t-distribution the power [Owen, 1965; Diletti, Hauschke, Steinijans, 1991] can be estimated from

$$1-\beta=\text{probt}\left(-t_{1-\alpha,n-2},n-2,\tau_2\right)-\text{probt}\left(t_{1-\alpha,n-2},n-2,\tau_1\right), \qquad (2.3.32)$$

where τ_1 and τ_2 are defined as

$$\tau_1=\frac{((\mu_A-\mu_B)+d)\sqrt{n}}{\sqrt{2\sigma_w^2}} \text{ and } \tau_2=\frac{((\mu_A-\mu_B)-d)\sqrt{n}}{\sqrt{2\sigma_w^2}}.$$

For quick calculations the sample size can estimate from

$$n=\frac{2\sigma_w^2\left(Z_{1-\beta}+Z_{1-\alpha}\right)^2}{((\mu_A-\mu_B)-d)^2}, \qquad (2.3.33)$$

and for very quick calculations, for 90% power and Type I error of 2.5%, one can use the following

$$n = \frac{21\sigma_w^2}{((\mu_A - \mu_B) - d)^2}. \qquad (2.3.34)$$

2.4.2.3. Special Case of No Treatment Difference

For the special case of $\mu_A - \mu_B = 0$ a direct estimate of the sample size can be estimated from

$$n = \frac{2\sigma_w^2(Z_{1-\beta/2} + Z_{1-\alpha})^2}{d^2}, \qquad (2.3.35)$$

which, if the variance is to be considered known for the statistical analysis, can be rewritten as

$$n = \frac{2\sigma_w^2(Z_{1-\beta/2} + t_{1-\alpha,n-2})^2}{d^2}. \qquad (2.3.36)$$

Equation (2.3.28) can be rewritten as

$$1-\beta = 2\Phi\left(\sqrt{\frac{d^2 n}{2\sigma_w^2}} - t_{1-\alpha,n-2}\right) - 1, \qquad (2.3.37)$$

which in turn, under the assumption of a non-central t-distribution, can also be rewritten as

$$1-\beta = 2\text{probt}(-t_{1-\alpha,n-2}, n-2, \tau) - 1, \qquad (2.3.38)$$

where τ is defined as

$$\tau = \frac{-\sqrt{n}d}{\sqrt{2\sigma_w^2}}.$$

For quick calculations, for 90% power and Type I error of 2.5%, the result formula can be used

$$n = \frac{26\sigma_w^2}{d^2}. \qquad (2.3.39)$$

As with parallel groups the quick equations give reasonable estimates of the sample size, underestimating the sample size by just one or two subjects, and thus provide reasonable initial values for iterations. Table 2.13 gives sample sizes using (2.3.32) for various standardised equivalence limits ($\delta = d/\sigma$) and mean differences.

Table 2-13. Total sample sizes (n) for cross-over equivalence study for different standardised equivalence limits and true mean differences (as a percentage of the equivalence limit) for 90% power and type I error rate of 2.5%

	Percentage Mean Difference				
d	0%	10%	15%	20%	25%
0.10	2601	2763	2981	3307	3742
0.20	652	692	747	828	937
0.30	291	309	333	370	418
0.40	165	175	189	209	236
0.50	106	113	122	135	152
0.60	75	79	85	94	106
0.70	56	59	63	70	79
0.80	43	46	49	54	61
0.90	35	37	39	43	49
1.00	29	30	32	36	40

2.4.2.4. Sensitivity Analysis About the Variance Used in the Sample Size Calculations

Table 2.14 gives the total sample size required for different standardised equivalence limits for a cross-over equivalence trial along with the sensitivity of these sample sizes to the 95th percentile of the variance (for different degrees of freedom) and different mean differences (assuming the mean difference is zero).

The inference and conclusions from the table are virtually identical to those for parallel group trials.

2.4.2.5. Calculations Taking Account of the Imprecision of the Variance Used in the Sample Size Calculations

2.4.2.6. General Case

To account for the degrees of freedom of the within subject sample variance the following equation could be used to calculate the power.

$$1-\beta = \text{probt}\left(-\tau_2, m, -t_{1-\alpha,n-2}\right) - \text{probt}\left(\tau_1, m, t_{1-\alpha,n-2}\right) \qquad (2.3.40)$$

where τ_1 and τ_2 are the absolute standardised equivalence limits defined as defined as

$$\tau_1 = \frac{|(\mu_A - \mu_B) - d|\sqrt{n}}{\sqrt{2}s_w} \text{ and } \tau_2 = \frac{|(\mu_A - \mu_B) + d|\sqrt{n}}{\sqrt{2}s_w}.$$

To calculate the sample size one needs to iterate to find the minimum value that would give the required power from (2.3.40).

Table 2-14. Total sample sizes for a cross-over equivalence trial for different standardised equivalence limits with 90% power and 2.5% type I error rate along with the power corresponding to the 95th percentile of the variance for different degrees of freedom and different mean differences

d	Sample Size	True Mean Diff (%)	Degrees of Freedom 10	25	50	75	100	500	∞
0.05	10398	0	0.24	0.57	0.70	0.75	0.78	0.85	0.90
		5	0.24	0.57	0.70	0.75	0.77	0.85	0.89
		10	0.23	0.56	0.68	0.73	0.76	0.83	0.88
		15	0.23	0.54	0.66	0.70	0.73	0.80	0.85
		20	0.22	0.51	0.62	0.67	0.69	0.76	0.81
		25	0.20	0.47	0.58	0.62	0.64	0.72	0.77
0.10	2601	0	0.24	0.57	0.70	0.75	0.78	0.85	0.90
		5	0.24	0.57	0.70	0.75	0.77	0.85	0.89
		10	0.23	0.56	0.68	0.73	0.76	0.83	0.88
		15	0.23	0.54	0.66	0.70	0.73	0.80	0.85
		20	0.22	0.51	0.62	0.67	0.69	0.76	0.81
		25	0.20	0.47	0.58	0.62	0.64	0.72	0.77
0.15	1158	0	0.24	0.57	0.70	0.75	0.78	0.86	0.90
		5	0.24	0.57	0.70	0.75	0.77	0.85	0.89
		10	0.23	0.56	0.68	0.73	0.76	0.83	0.88
		15	0.23	0.54	0.66	0.70	0.73	0.80	0.85
		20	0.22	0.51	0.62	0.67	0.69	0.76	0.81
		25	0.20	0.47	0.58	0.62	0.64	0.72	0.77
0.20	652	0	0.24	0.57	0.70	0.75	0.78	0.86	0.90
		5	0.24	0.57	0.70	0.75	0.77	0.85	0.89
		10	0.23	0.56	0.68	0.73	0.76	0.83	0.88
		15	0.23	0.54	0.66	0.70	0.73	0.80	0.85
		20	0.22	0.51	0.62	0.67	0.69	0.76	0.81
		25	0.20	0.47	0.58	0.62	0.64	0.72	0.77
0.25	418	0	0.24	0.57	0.70	0.75	0.78	0.86	0.90
		5	0.24	0.57	0.70	0.75	0.77	0.85	0.89
		10	0.23	0.56	0.68	0.73	0.76	0.83	0.88
		15	0.23	0.54	0.66	0.70	0.73	0.80	0.85
		20	0.22	0.51	0.62	0.67	0.69	0.76	0.81
		25	0.20	0.47	0.58	0.62	0.64	0.72	0.77

For non-zero treatment differences (i.e. for $\mu_A - \mu_B > 0$) the power could be estimated from

$$1 - \beta = 1 - \text{probt}\left(\frac{|(\mu_A - \mu_B) - d|\sqrt{n}}{\sqrt{2}s_w}, m, t_{1-\alpha, n-2} \right). \qquad (2.3.41)$$

Which when written in terms of n becomes

$$n \geq \frac{2s_w^2\left[tinv(1-\beta, \mathrm{m}, t_{1-\alpha, n-2})\right]^2}{((\mu_A - \mu_B) - d)^2}. \qquad (2.3.42)$$

Replacing the t-statistic with a Z-statistic and (2.3.42) can in turn be approximated from

$$n = \frac{2s_w^2\left[tinv(1-\beta, \mathrm{m}, Z_{1-\alpha})\right]^2}{\left((\mu_A - \mu_B) - d\right)^2}. \qquad (2.3.43)$$

This direct estimate could be used to provide initial estimates of the sample size for (2.3.40).

2.4.2.7. Special Case of No Treatment Difference

For the special case of no treatment difference the power can be estimate from

$$1-\beta = 2\mathrm{probt}\left(\tau, m, -t_{1-\alpha, n-2}\right) - 1, \qquad (2.3.44)$$

where τ is defined as

$$\tau = \frac{-\sqrt{n}d}{\sqrt{2}s_w},$$

Which when written in terms of n becomes

$$n \geq \frac{2s_w^2\left[tinv(1-\beta\ 2, \mathrm{m}, t_{1-\alpha, n-2})\right]^2}{d^2}. \qquad (2.3.45)$$

Replacing the t-statistic with a Z-statistic and (2.3.45) can in turn be approximated from

$$n = \frac{2s_w^2\left[tinv(1-\beta\ 2, \mathrm{m}, Z_{1-\alpha})\right]^2}{d^2}. \qquad (2.3.46)$$

Table 2.15 gives the sample sizes for different degrees of freedom and standardised equivalence limits. For cross equivalence trials, as with superiority trials, the multiplication factors only depend on the Type I and II errors and the degrees of freedom and are the same as for parallel group trials given in Table 2.12

Table 2-15. Sample sizes estimated for different standardised equivalence limits and degrees of freedom from (2.3.32). The final column with "infinite" degrees of freedom is from (2.3.41) and the assumption that the within subject population variance is being used. The type I error is set at a two one-sided significance level of 2.5% and the type II error is set at 10%

	Degrees of Freedom						
d	10	25	50	75	100	500	∞
0.10	3706	2991	2788	2724	2693	2619	2601
0.20	928	749	698	682	675	656	652
0.30	414	334	311	304	301	293	291
0.40	233	189	176	172	170	165	164
0.50	150	121	113	111	109	107	106
0.60	105	85	79	77	77	75	74
0.65	90	73	68	66	66	64	63
0.70	78	63	59	57	57	55	55
0.80	60	49	45	44	44	43	42
0.90	48	39	36	36	35	34	34
1.00	39	32	30	29	29	28	28

2.4.2.8. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations

For the general case of $\bar{x}_A - \bar{x}_B \neq 0$ the power for a given sample size can be estimated (through numerical integration) from

$$1-\beta = \sum_{p=0.001}^{0.998} 0.001x \left(\left[\begin{array}{l} \text{probt}\left(\dfrac{\sqrt{n}\left((\bar{x}_A - \bar{x}_B) + Z_p se(\bar{x}_A - \bar{x}_B) - d\right)}{\sqrt{2s_w^2}}, m, -t_{1-\alpha,n-2} \right) \\ - \text{probt}\left(\dfrac{\sqrt{n}\left((\bar{x}_A - \bar{x}_B) + Z_p se(\bar{x}_A - \bar{x}_B) + d\right)}{\sqrt{2s_w^2}}, m, t_{1-\alpha,n-2} \right) \end{array} \right] + \left[\begin{array}{l} \text{probt}\left(\dfrac{\sqrt{n}\left((\bar{x}_A - \bar{x}_B) + Z_{p+0.001} se(\bar{x}_A - \bar{x}_B) - d\right)}{\sqrt{2s_w^2}}, m, -t_{1-\alpha,n-2} \right) \\ - \text{probt}\left(\dfrac{\sqrt{n}\left((\bar{x}_A - \bar{x}_B) + Z_{p+0.001} se(\bar{x}_A - \bar{x}_B) + d\right)}{\sqrt{2s_w^2}}, m, t_{1-\alpha,n-2} \right) \end{array} \right] \right) / 2 \quad , \qquad (2.3.47)$$

where the sample size required is the minimum value which gives the required sample size. Z_p is the value from the Normal distribution that equates to the percentile p. As with parallel group trials there is little practical application of this result.

2.4.2.9. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations – A Bayesian Approach

In this section the Bayesian methods described for parallel group data will be extended to cross-over data.

2.4.2.10. Prior Response

Prior values for the mean difference and within subject population standard deviation are defined as d_0 (i.e. for $\mu_A - \mu_B$) and s_{w0} respectively. As with parallel group data these values can be subjective values taken as beliefs about the mean difference.

2.4.2.11. Anticipated Response

The anticipated mean difference and within subject population standard deviation are defined as d_1 and s_{w1} respectively. These values are taken as objective values observed in a previous clinical trial. Hence $\sqrt{2}s_{w1}/n$ is an estimate of the within subject standard deviation around the mean.

2.4.2.12. Posterior Response

With the anticipated and prior responses the posterior distribution can be calculated through a weighted sum of the prior and anticipated responses. The posterior estimate of the within subject population variance around the mean, s^2_{wn}, is defined as

$$s^2_{wn} = \left(\frac{n}{2s^2_{w1}} + \frac{1}{s^2_{w0}} \right)^{-1}, \tag{2.3.48}$$

and the posterior estimate of the mean difference, d_n, is defined as

$$d_n = s^2_{wn} \left(\frac{d_0}{s^2_{w0}} + \frac{d_1 n}{2s^2_{w1}} \right). \tag{2.3.49}$$

From these posterior values a density distribution for $prob(\theta < d_i \mid d_1)$ can be defined such that a probability of observing d_i, or greater, for a given d_n would be

$$prob(\theta > d_i \mid d_1) = \Phi\left(\frac{d_n - d_i}{s_{wn}} \right), \tag{2.3.50}$$

or alternatively for $prob(\theta < d_i \mid d_1)$

$$prob(\theta > d_i \mid d_1) = \Phi\left(\frac{d_i - d_n}{s_n} \right).$$

From this result a mean difference d for given percentile p can be estimated and put into the following result for the general case $\bar{x}_A - \bar{x}_B \neq 0$, to estimate the power for a given sample size

$$1 - \beta = \sum_{p=0.001}^{0.998} 0.001x \left(\left[\text{probt}\left(\frac{\sqrt{n}(d_p - d)}{\sqrt{2}s_w}, m, -t_{1-\alpha, n-2} \right) + \text{probt}\left(\frac{\sqrt{n}(d_p + d)}{\sqrt{2}s_w}, m, t_{1-\alpha, n-2} \right) \right] + \left[\text{probt}\left(\frac{\sqrt{n}(d_{p+0.001} - d)}{\sqrt{2}s_w}, m, -t_{1-\alpha, n-2} \right) + \text{probt}\left(\frac{\sqrt{n}(d_{p+0.001} + d)}{\sqrt{2}s_w}, m, t_{1-\alpha, n-2} \right) - 1 \right] \right) / 2, \tag{2.3.51}$$

where the sample size is estimated through numerical integration and iteration.

2.5. Non-Inferiority Trials

2.5.1. Parallel Group Trials

2.5.1.1. Sample Size Estimated Assuming the Population Variance to be Known

From Chapter 1 one requires

$$Var\ (S) = \frac{(d-\Delta)^2}{(Z_{1-\alpha} + Z_{1-\beta})^2}, \tag{2.4.1}$$

and as with superiority and equivalence trials Var(S) can be defined as

$$Var\ (S) = \frac{r+1}{r} \cdot \frac{\sigma^2}{n_A},$$

which can be substituted in to (2.4.1) (replacing Δ with $\mu_A - \mu_B$) giving a direct estimate of the sample size

$$n_A = \frac{(r+1)\sigma^2 (Z_{1-\beta} + Z_{1-\alpha})^2}{r((\mu_A - \mu_B) - d)^2}. \tag{2.4.2}$$

Rewriting (2.4.2) to give power for a give sample size results in

$$1-\beta = \Phi\left(\sqrt{\frac{((\mu_A - \mu_B) - d)^2 r n_A}{(r+1)\sigma^2}} - Z_{1-\alpha} \right).$$

The equivalent for the case when the variance is considered unknown for the analysis is

$$1-\beta = \Phi\left(\sqrt{\frac{((\mu_A - \mu_B) - d)^2 r n_A}{(r+1)\sigma^2}} - t_{1-\alpha, n_A(r+1)-2} \right). \tag{2.4.3}$$

As with the sections on equivalence and superiority trials it is best to calculate the power under the assumption of a non-central t-distribution [Julious, 2004a]

$$1-\beta = 1 - \text{probt}(t_{1-\alpha, n_A(r+1)-2}, n_A(r+1) - 2, \tau), \tag{2.4.4}$$

where τ is defined as

$$\tau = \left| \frac{((\mu_A - \mu_B) - d)\sqrt{r n_A}}{\sqrt{(r+1)\sigma^2}} \right|.$$

For quick calculations, for 90% power and Type I error of 2.5%, the following formula can be used

$$n_A = \frac{10.5\sigma^2(r+1)}{((\mu_A - \mu_B) - d)^2 r}. \qquad (2.4.5)$$

In the case of r=1 (2.4.5) resolves to

$$n_A = \frac{21\sigma^2}{((\mu_A - \mu_B) - d)^2}. \qquad (2.4.6)$$

The quick equations give reasonable estimates of the sample size, although with slight underestimation. Table 2.16 gives sample sizes using (2.4.4) for various standardised non-inferiority limits ($\delta = d/\sigma$) and standardised mean differences assuming equal allocation between groups.

Table 2-16. Sample sizes (n_A) for one arm of a parallel group non-Inferiority study with equal allocation for different standardised non-inferiority limits and true mean differences (as a percentage of the non-inferiority limit) for 90% power and type I error rate of 2.5%

	Percentage Mean Difference										
d	-25%	-20%	-15%	-10%	-5%	0%	5%	10%	15%	20%	25%
0.10	1346	1461	1590	1738	1908	2103	2330	2596	2910	3285	3737
0.20	338	366	399	436	478	527	584	650	729	822	935
0.30	151	164	178	194	213	235	260	290	325	366	417
0.40	86	93	101	110	121	133	147	164	183	207	235
0.50	55	60	65	71	78	86	95	105	118	133	151
0.60	39	42	46	50	54	60	66	74	82	93	105
0.70	29	31	34	37	40	44	49	54	61	68	78
0.80	23	24	26	29	31	34	38	42	47	53	60
0.90	18	20	21	23	25	27	30	34	37	42	48
1.00	15	16	17	19	21	23	25	27	31	34	39

One feature to highlight in Tables 2.16 and 2.20 (described in the next section on cross-over trials) is the asymmetric effect on the sample size for different values of the true mean difference. In equivalence trials as one has two, usually symmetric, margins, when one moves away from a zero mean difference in any direction the sample size is inflated. However, in non-inferiority trials the sample size is inflated only if the true mean difference moves towards the non-inferiority margin. If it is expected that the true mean difference is in favour of the comparator regimen (compared to control) then the sample size is significantly reduced.

This asymmetric effect of the mean difference on the sample size should be considered when designing non-inferiority trials as even only a small expected mean difference in favour of the comparator could have a marked effect on the sample size.

2.5.1.2. Worked Example

An investigator wishes to design a hypertension trial where the objective is to demonstrate that one treatment (an investigative therapy) is non-inferior to another (a standard therapy). The largest clinically acceptable effect to be able to declare non-inferiority is a change in blood pressure of 5mmHg (d). The true mean difference between the treatments is thought to be zero with an expected standard deviation in the trial population of 25mmHg (σ). There is to be equal allocation between groups. Thus, the standardised non-inferiority limits equate to $-\delta = -d/\sigma = -5/25 = -0.20$. For the Type I and Type II errors fixed at 2.5% and 10% respectively Table 2.16 gives a sample size of 527 patients in each arm of the trial.

Suppose, though, that one believes that the investigative therapy is a little superior to the standard such that the true mean difference is thought to be 1mmHg. This inflates the distance one expects the mean to be away from the non-inferiority margin by 20% and would reduce the sample size required to 366 patients in each arm of the trial.

2.5.1.3. Sensitivity Analysis About the Variance Used in the Sample Size Calculations

As with superiority and equivalence trials described earlier (2.2.9) can be used to estimate a plausibly large value for the population variance. As with equivalence trials however, one also needs to investigate the sensitivity of calculations to the assumption about the true mean difference. If one has assumed no difference when it was truly non-zero then this will have an effect on the power of the study. Unlike equivalence trials the adverse affect on the power is not symmetric and if there is a difference in favour of the investigative treatment there will be a positive effect on the power. This section will only investigate negative effects.

Table 2.17 gives the sample size per group required for different standardised non-inferiority limits for a parallel group equivalence trial along with the sensitivity of these sample sizes to the 95th percentile of the variance (for different degrees of freedom) and different mean differences (assuming the mean difference is zero).

From Table 2.17 it seems that assuming a trial is designed with 90% power:

- That to ensure that the study has at least 50% or 80% power, if the true variance is nearer the 95th percentile, then the variance estimate would require at least 10 and 75 degrees of freedom respectively – assuming that there was no mean difference.

- If one's variance estimate was close to the true variance then even if the true mean difference was 15% of the standardised non-inferiority limit (against the investigative treatment) then one would still have 80% power in the study.

Table 2-17. Sample sizes per arm for a parallel group non-inferiority trial for different standardised non-inferiority limits with 90% power and 2.5% type I error rate along with the power corresponding to the 95th percentile of the variance for different degrees of freedom and different mean differences

		True Mean Diff	Degrees of Freedom						
d	Sample Size	(%)	10	25	50	75	100	500	∞
0.05	8407	0	0.53	0.70	0.77	0.80	0.82	0.87	0.90
		5	0.49	0.65	0.73	0.76	0.78	0.83	0.87
		10	0.45	0.61	0.68	0.71	0.73	0.79	0.83
		15	0.41	0.56	0.63	0.66	0.68	0.74	0.79
		20	0.37	0.51	0.58	0.61	0.63	0.69	0.74
		25	0.33	0.46	0.53	0.56	0.57	0.63	0.68
0.10	2103	0	0.53	0.70	0.77	0.80	0.82	0.87	0.90
		5	0.49	0.65	0.73	0.76	0.78	0.83	0.87
		10	0.45	0.61	0.68	0.71	0.73	0.79	0.83
		15	0.41	0.56	0.63	0.66	0.68	0.74	0.79
		20	0.37	0.51	0.58	0.61	0.63	0.69	0.74
		25	0.33	0.46	0.53	0.56	0.57	0.63	0.68
0.15	935	0	0.53	0.70	0.77	0.80	0.82	0.87	0.90
		5	0.49	0.65	0.73	0.76	0.78	0.83	0.87
		10	0.45	0.61	0.68	0.71	0.73	0.79	0.83
		15	0.41	0.56	0.63	0.66	0.68	0.74	0.79
		20	0.37	0.51	0.58	0.61	0.63	0.69	0.74
		25	0.33	0.46	0.53	0.56	0.57	0.63	0.68
0.20	527	0	0.53	0.70	0.77	0.80	0.82	0.87	0.90
		5	0.49	0.65	0.73	0.76	0.78	0.83	0.87
		10	0.45	0.61	0.68	0.71	0.73	0.79	0.83
		15	0.41	0.56	0.63	0.66	0.68	0.74	0.79
		20	0.37	0.51	0.58	0.61	0.63	0.69	0.74
		25	0.33	0.46	0.53	0.56	0.57	0.64	0.68
0.25	338	0	0.53	0.70	0.77	0.80	0.82	0.87	0.90
		5	0.49	0.65	0.73	0.76	0.78	0.83	0.87
		10	0.45	0.61	0.68	0.71	0.73	0.79	0.83
		15	0.41	0.56	0.63	0.66	0.68	0.74	0.79
		20	0.37	0.51	0.58	0.61	0.63	0.69	0.74
		25	0.33	0.46	0.53	0.56	0.57	0.64	0.68

2.5.1.4. Worked Example

Revisiting the worked example earlier. Suppose the sample variance used in calculations was estimated with 10 degrees of freedom. Table 2.17 demonstrates that a high plausible value for this variance would have 53% power with the same sample size. If the true mean difference was 10% of the non-inferiority limit, against the investigative treatment, the power of the study would be reduced to 83%.

2.5.1.5. Calculations Taking Account of the Imprecision of the Variance Used in the Sample Size Calculations

To account for the imprecision of the sample variance used in the sample size calculations the results given in the section on superiority trials can be generalised to give the following result

$$n \geq \frac{(r+1)s^2\left[tinv\left(1-\beta, \mathrm{m}, t_{1-\alpha, n_A(r+1)-2}\right)\right]^2}{r\left((\mu_A-\mu_B)-d\right)^2}, \qquad (2.4.7)$$

where the sample size required is the least integer value for (2.4.7) to hold. One can rewrite (2.4.7) in terms of power to obtain the following result

$$1-\beta = 1-\mathrm{probt}\left(\tau, m, t_{1-\alpha, n_A(r+1)-2}\right), \qquad (2.4.8)$$

where τ is defined as

$$\tau = \left|\frac{\left((\mu_A-\mu_B)-d\right)\sqrt{rn_A}}{\sqrt{(r+1)s^2}}\right|.$$

Replacing the t-statistic with a Z-statistic gives one the following result

$$n = \frac{(r+1)s^2\left[tinv\left(1-\beta, \mathrm{m}, Z_{1-\alpha/2}\right)\right]^2}{r\left((\mu_A-\mu_B)-d\right)^2}, \qquad (2.4.9)$$

which allows one to have a direct estimate of the sample size and also gives an initial value for iterations for (2.4.7). Tables 2.18 and 2.19 are produced for the special case of no mean difference between treatments.

Table 2-18. Sample sizes estimated for different standardised non-inferiority limits and degrees of freedom from (2.4.7). The final column with "infinite" degrees of freedom is from (2.4.3) and the assumption that the population variance is being used. The type I error is set at a one-sided significance levels of 2.5% and the type II error is set at 10%

	Degrees of Freedom						
D	10	25	50	75	100	500	∞
0.10	2734	2333	2214	2176	2158	2114	2103
0.20	685	584	555	545	540	529	527
0.30	305	260	247	243	241	236	235
0.40	172	147	140	137	136	133	133
0.50	111	95	90	88	88	86	85
0.60	77	66	63	62	61	60	60
0.70	57	49	46	46	45	44	44
0.80	44	38	36	35	35	34	34
0.90	35	30	29	28	28	27	27
1.00	29	25	23	23	23	22	22

Table 2.18 gives the sample sizes for different degrees of freedom and standardised equivalence limits. Table 2.19 gives the multiplication factors, compared to assuming one has the population variance, for various degrees of freedom and Type I and II errors. These multiplication factors can be used to inflate a sample size to account for the imprecision in the variance.

Table 2-19. Multiplication factors for different levels of one sided significance, type II error and degrees of freedom

		Significance Level (α)			
m	β	0.010	0.025	0.050	0.100
5	0.05	2.167	2.068	1.980	1.875
	0.10	1.776	1.711	1.652	1.581
	0.15	1.582	1.533	1.489	1.436
	0.20	1.457	1.419	1.385	1.344
	0.50	1.120	1.117	1.114	1.111
10	0.05	1.463	1.425	1.392	1.353
	0.10	1.328	1.301	1.276	1.248
	0.15	1.254	1.233	1.214	1.192
	0.20	1.204	1.187	1.172	1.154
	0.50	1.055	1.054	1.053	1.053
25	0.05	1.163	1.150	1.139	1.125
	0.10	1.119	1.109	1.101	1.091
	0.15	1.094	1.086	1.079	1.071
	0.20	1.076	1.070	1.065	1.058
	0.50	1.021	1.021	1.021	1.020
50	0.05	1.078	1.072	1.067	1.060
	0.10	1.058	1.053	1.049	1.044
	0.15	1.046	1.042	1.039	1.035
	0.20	1.037	1.034	1.032	1.028
	0.50	1.010	1.010	1.010	1.010
75	0.05	1.052	1.047	1.044	1.040
	0.10	1.038	1.035	1.032	1.029
	0.15	1.030	1.028	1.026	1.023
	0.20	1.025	1.023	1.021	1.019
	0.50	1.007	1.007	1.007	1.007
100	0.05	1.038	1.035	1.033	1.030
	0.10	1.029	1.026	1.024	1.022
	0.15	1.023	1.021	1.019	1.017
	0.20	1.019	1.017	1.016	1.014
	0.50	1.005	1.005	1.005	1.005

2.5.1.6. Worked Example

Returning to the worked example given earlier for a standardised non-inferiority limit of 0.20 with 10 degrees of freedom for the sample variance estimate Table 2.18 gives a sample size of 685. This compares to sample size of 527 calculated assuming one had

the population variance for calculations - a potential under estimation of the sample size of 30%

2.5.1.7. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations

Similarly to equivalence trials as well as being sensitive to the assumptions around the variance the sample size calculations are also sensitive to the assumptions around the assumed mean difference. As the mean difference deviates away from the assumptions in the calculations there is a consequent effect on the power. However, this effect is not symmetric. If one had underestimated an effect in favour of the investigative treatment then one may have a study that is overpowered. On the other hand if one had underestimated an effect against the investigative treatment then one may have an underpowered study.

With an estimate of the mean difference ($\bar{x}_A - \bar{x}_B$) from a previous study and an estimate of the standard error around this mean difference ($se(\bar{x}_A - \bar{x}_B)$) an estimate of the sample size can be obtained. For the special case $\bar{x}_A - \bar{x}_B \neq 0$ the power for a given sample size can be estimated (through numerical integration) from

$$1-\beta = \sum_{p=0.001}^{0.998} 0.001x\left[\begin{array}{l} 2 - \text{probt}\left(\dfrac{\sqrt{rn_A}\left|(\bar{x}_A - \bar{x}_B) + Z_p se(\bar{x}_A - \bar{x}_B) - d\right|}{\sqrt{(r+1)s}}, m, t_{1-\alpha, n_A(r+1)-2}\right) \\ - \text{probt}\left(\dfrac{\sqrt{rn_A}\left|(\bar{x}_A - \bar{x}_B) + Z_{p=0.001} se(\bar{x}_A - \bar{x}_B) - d\right|}{\sqrt{(r+1)s}}, m, t_{1-\alpha, n_A(r+1)-2}\right) \end{array}\right] / 2, \qquad (2.4.10)$$

where the sample size required is the minimum value which gives the required sample size. As with equivalence trials discussed earlier this result however is practically unappealing.

2.5.1.8. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations – A Bayesian Approach

The prior, anticipated and posterior responses would be calculated as per parallel group equivalence trials. Hence, a mean difference d_i for given percentile p can be estimated and put into the following result for the general case $\bar{x}_A - \bar{x}_B \neq 0$, to estimate the power for a given sample size.

$$1-\beta = \sum_{p=0.001}^{0.998} 0.001x\left(2 - \text{probt}\left(\frac{\sqrt{rn_A}\left|d_p - d\right|}{\sqrt{(r+1)s^2}}, m, t_{1-\alpha, n_A(r+1)-2}\right) - \text{probt}\left(\frac{\sqrt{rn_A}\left|d_{p=0.001} - d\right|}{\sqrt{(r+1)s^2}}, m, t_{1-\alpha, n_A(r+1)-2}\right)\right) / 2, \qquad (2.4.12)$$

where the sample size is estimated through numerical integration and iteration.

2.5.2. Cross-over Trials

2.5.2.1. Sample Size Estimated Assuming the Population Variance to be Known

The equivalent sample size formula to (2.4.2) for cross-over trials is

$$n = \frac{2\sigma_w^2\left(Z_{1-\beta} + Z_{1-\alpha}\right)^2}{\left((\mu_A - \mu_B) - d\right)^2}, \tag{2.4.12}$$

which when rewritten in terms of power becomes

$$1 - \beta = \Phi\left(\sqrt{\frac{\left((\mu_A - \mu_B) - d\right)^2 n}{2\sigma_w^2}} - Z_{1-\alpha}\right). \tag{2.4.13}$$

The equivalent formula replacing the Z-statistics with the t-statistic is

$$1 - \beta = \Phi\left(\sqrt{\frac{\left((\mu_A - \mu_B) - d\right)^2 n}{2\sigma_w^2}} - t_{1-\alpha,n-2}\right). \tag{2.4.14}$$

As with parallel group designs it preferable to calculate the Type II error (and power) under the assumption of a non-central t-distribution and thus (2.4.14) is rewritten as [Julious 2004a]

$$1 - \beta = 1 - \text{probt}\left(t_{1-\alpha,n-2}, n-2, \tau\right), \tag{2.4.15}$$

where τ is defined as

$$\tau = \left|\frac{\left((\mu_A - \mu_B) - d\right)\sqrt{n}}{\sqrt{2\sigma_w^2}}\right|.$$

For quick calculations, for 90% power and Type I error of 2.5%, the following formula can be utilised

$$n = \frac{21\sigma_w^2}{\left((\mu_A - \mu_B) - d\right)^2}. \tag{2.4.16}$$

As with parallel group sample size estimation the quick equations give reasonable, although slightly under, estimates of the sample size. Table 2.20 gives sample sizes using (2.4.15) for various standardised equivalence limits ($\delta = d/\sigma$) and standardised mean differences assuming equal allocation between groups.

Table 2-20. Total sample sizes (n) for a cross-over non-inferiority study with equal allocation for different standardised non-inferiority limits and true mean differences (as a percentage of the equivalence limit) for 90% power and type I error rate of 2.5%

	Percentage Mean Difference										
d	-25%	-20%	-15%	-10%	-5%	0%	5%	10%	15%	20%	25%
0.10	1347	1462	1591	1739	1909	2104	2331	2597	2911	3286	3738
0.20	339	367	400	437	479	528	585	651	730	823	936
0.30	152	165	179	195	214	236	261	291	326	367	418
0.40	87	94	102	111	122	134	148	165	184	208	236
0.50	56	61	66	72	79	87	96	106	119	134	152
0.60	40	43	47	51	55	61	67	75	83	94	106
0.70	30	32	35	38	41	45	50	55	62	69	79
0.80	24	25	27	30	32	35	39	43	48	54	61
0.90	19	21	22	24	26	29	31	35	38	43	49
1.00	16	17	19	20	22	24	26	29	32	35	40

2.5.2.2. Sensitivity Analysis About the Variance Used in the Sample Size Calculations

Table 2.21 gives the total sample size required for different standardised non-inferiority for a cross-over trial along with the sensitivity of these sample sizes to the 95th percentile of the variance (for different degrees of freedom) and different mean differences (assuming the mean difference is zero). The inference and conclusions from the table are virtually identical to those for parallel group trials.

2.5.2.3. Calculations Taking Account of the Imprecision of the Variance Used in the Sample Size Calculations

To account for the imprecision of the variance used in the sample size calculations the results for parallel group trials can be generalised to

$$n \geq \frac{s_w^2\left[tinv(1-\beta, \mathrm{m}, t_{1-\alpha,n-2})\right]^2}{[(\mu_A-\mu_B)-d]^2}, \qquad (2.4.17)$$

where n is the least integer value for (2.4.17) to hold. One can rewrite (2.4.17) in terms of power

$$1-\beta = 1-probt\left(\frac{\sqrt{n}|(\mu_A-\mu_B)-d|}{\sqrt{2s_w^2}}, m, t_{1-\alpha,n-2}\right). \qquad (2.4.18)$$

Replacing the t-statistic with a Z-statistic gives one the following result

$$n = \frac{2s_w^2\left[tinv(1-\beta, \mathrm{m}, Z_{1-\alpha/2})\right]^2}{[(\mu_A-\mu_B)-d]^2} \qquad (2.4.19)$$

which allows one to have a direct estimate of the sample size and also gives an initial value for (2.2.17).

Table 2-21. Total sample sizes for a cross-over non-inferiority trial for different standardised non-inferiority limits with 90% power and 2.5% type I error rate along with the power corresponding to the 95th percentile of the variance for different degrees of freedom and different mean differences

		True Mean							
	Sample	Diff	Degrees of Freedom						
d	Size	(%)	10	25	50	75	100	500	∞
0.05	8408	0	0.53	0.70	0.77	0.80	0.82	0.87	0.90
		5	0.49	0.65	0.73	0.76	0.78	0.83	0.87
		10	0.45	0.61	0.68	0.71	0.73	0.79	0.83
		15	0.41	0.56	0.63	0.66	0.68	0.74	0.79
		20	0.37	0.51	0.58	0.61	0.63	0.69	0.74
		25	0.33	0.46	0.53	0.56	0.57	0.63	0.68
0.10	2104	0	0.53	0.70	0.77	0.80	0.82	0.87	0.90
		5	0.49	0.65	0.73	0.76	0.78	0.83	0.87
		10	0.45	0.61	0.68	0.71	0.73	0.79	0.83
		15	0.41	0.56	0.63	0.66	0.68	0.74	0.79
		20	0.37	0.51	0.58	0.61	0.63	0.69	0.74
		25	0.33	0.46	0.53	0.56	0.57	0.63	0.68
0.15	936	0	0.53	0.70	0.77	0.80	0.82	0.87	0.90
		5	0.49	0.65	0.73	0.76	0.78	0.83	0.87
		10	0.45	0.61	0.68	0.71	0.73	0.79	0.83
		15	0.41	0.56	0.63	0.66	0.68	0.74	0.79
		20	0.37	0.51	0.58	0.61	0.63	0.69	0.74
		25	0.33	0.46	0.53	0.56	0.57	0.63	0.68
0.20	528	0	0.53	0.70	0.77	0.80	0.82	0.87	0.90
		5	0.49	0.65	0.73	0.76	0.78	0.83	0.87
		10	0.45	0.61	0.68	0.71	0.73	0.79	0.83
		15	0.41	0.56	0.63	0.66	0.68	0.74	0.79
		20	0.37	0.51	0.58	0.61	0.63	0.69	0.74
		25	0.33	0.46	0.53	0.56	0.57	0.64	0.68
0.25	339	0	0.53	0.70	0.77	0.80	0.82	0.87	0.90
		5	0.49	0.65	0.73	0.76	0.78	0.83	0.87
		10	0.45	0.61	0.68	0.71	0.73	0.79	0.83
		15	0.41	0.56	0.63	0.66	0.68	0.74	0.79
		20	0.37	0.51	0.58	0.61	0.63	0.69	0.74
		25	0.33	0.46	0.53	0.56	0.57	0.64	0.68

Table 2.22 give the sample sizes required using (2.4.17) for different standardised non-inferiority limits and degrees of freedom. For multiplication factors see Table 2.19. As these factors depend only on the Type I error, Type II error and degrees of freedom they are the same for both parallel and cross-over trials.

Table 2-22. Sample sizes estimated for different standardised non-inferiority limits and degrees of freedom from (2.4.17). The final column with "infinite" degrees of freedom is from (2.4.14) and the assumption that the population variance is being used. The type I error is set at a one-sided significance level of 2.5% and the type II error is set at 10%

	Degrees of Freedom						
d	10	25	50	75	100	500	∞
0.10	2734	2333	2214	2176	2158	2114	2103
0.20	685	584	555	545	540	529	527
0.30	305	260	247	243	241	236	235
0.40	172	147	140	137	136	133	133
0.50	111	95	90	88	88	86	85
0.60	77	66	63	62	61	60	60
0.70	57	49	46	46	45	44	44
0.80	44	38	36	35	35	34	34
0.90	35	30	29	28	28	27	27
1.00	29	25	23	23	23	22	22

2.5.2.4. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations

Similarly to parallel group trials with an estimate of the mean difference ($\bar{x}_A - \bar{x}_B$) from a previous study and an estimate of the standard error around this mean difference ($se(\bar{x}_A - \bar{x}_B)$) an estimate of the sample size can be obtained. For the general case of $\bar{x}_A - \bar{x}_B \neq 0$ the power for a given sample size can be estimated (through numerical integration) from

$$1-\beta = \sum_{p=0.001}^{0.998} 0.001x \left[\begin{array}{l} 2 - \text{probt}\left(\dfrac{\sqrt{n}\left(\left|\bar{x}_A - \bar{x}_B\right| + Z_p se(\bar{x}_A - \bar{x}_B) - d\right|}{\sqrt{2}s_w}, m, t_{1-\alpha,n-2} \right) \\ - \text{probt}\left(\dfrac{\sqrt{n}\left(\left|\bar{x}_A - \bar{x}_B\right| + Z_{p+0.001} se(\bar{x}_A - \bar{x}_B) - d\right|}{\sqrt{2}s_w}, m, t_{1-\alpha,n-2} \right) - 2 \end{array} \right] / 2, \qquad (2.4.19)$$

where the sample size required is the minimum value which gives the required sample size.

2.5.2.5. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations – A Bayesian Approach

As with parallel group non-inferiority trials described earlier the priors, anticipated and posterior responses are defined as for equivalence trials. Hence, the power, for given n, can be calculated from

$$1-\beta = \sum_{p=0.001}^{0.998} 0.001x \left(2 - \text{probt}\left(\frac{\sqrt{n}\left|d_p - d\right|}{\sqrt{2}s}, m, t_{1-\alpha,n-2} \right) - \text{probt}\left(\frac{\sqrt{n}\left|d_{p+0.001} - d\right|}{\sqrt{2}s}, m, t_{1-\alpha,n-2} \right) \right) / 2 \qquad (2.4.20)$$

where the sample size is estimated through numerical integration and iteration.

2.6. As Good as or Better Trials

To calculate the sample size required for an "as good as or better" trial one should apply the methodologies described in Sections 2.2 (Superiority) and 2.4 (Non-inferiority). For example a parallel group trial to investigate a one sided test of non-inferiority and a two sided test of superiority; designed to about a standardised clinically meaningful difference for superiority of 0.20 and a standardised non inferiority margin of 0.20. With the Type I error fixed at 5% for the test of superiority and 2.5% for the test of non-inferiority and the Type II error fixed at 10%. From Table 2.1 for superiority one would require 527 patients in each arm. Whilst from Table 2.16 for non-inferiority, assuming no treatment difference, again one would require 527 patients per arm.

Note that here one is making the big, probably unrealistic, assumption that the standardised non-inferiority limit and the standardised difference are the same.

On the face of it then one can switch between non-inferiority and superiority whilst maintaining the Type I error for no great cost in the sample size. However, if in the example above, to test non-inferiority one wished to allow for the fact that there may be a true mean difference between the two groups against the investigative therapy. If this mean difference equated to 20% of the standardised non-inferiority limit it would inflate the sample size, mutatis mutardis, to 822 patients per arm.

A more realistic scenario to the one described in the previous paragraph is one where the non-inferiority margin is a fraction of the clinically meaningful difference. The sample size required to investigate non-inferiority would hence be a factor more than required to investigate superiority - the factor being the ratio of the clinically meaningful difference over the non-inferiority margin squared. In this circumstance, given that one also is investigating superiority, it may be appropriate to power for non-inferiority assuming a small difference between the two groups in favour of the investigative therapy.

A further consideration in as good as or better trials is the choice of data set to have as primary – which adds a further complication. For a superiority trial the primary data set would be that based on intention to treat (ITT); for a non-inferiority trial the primary data set would be both the per protocol data set (PP) and the ITT [CPMP, 2000].

2.7. Bioequivalence Trials

In a reversal to the ordering of previous sections the calculations for cross-over trials will be described before parallel group trials. The reason for this is that cross-over trials are the most common designs for bioequivalence studies.

2.7.1. Cross-over Trials

2.7.1.1. Sample Sizes Estimated Assuming the Population Variance to be Known

2.7.1.2. General Case

The derivation of the sample size equations is similar to that for equivalence trials. For the general case where the expected true mean difference is not fixed to be unity the sample size cannot be directly derived. One instead has to iterate until a sample size is reached which gives the required Type II error and power.

To calculate the power for the two one-sided test procedure at the 5% significance level where the bioequivalence acceptance limits are (0.80, 1.25) for any given value for the true ratio, μ_T/μ_R, the following formula can be used

$$1-\beta = \Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(1.25))^2 n}{2\sigma_w^2}} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(0.80))^2 n}{2\sigma_w^2}} - Z_{1-\alpha}\right) - 1, \quad (2.6.1)$$

where σ_w is the within-subject variability on the log scale and n is the total sample size. Replacing the Z-statistic with the t-statistic and (2.6.1) can be rewritten as

$$1-\beta = \Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(1.25))^2 n}{2\sigma_w^2}} - t_{1-\alpha,n-2}\right) + \Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(0.80))^2 n}{2\sigma_w^2}} - t_{1-\alpha,n-2}\right) - 1. \quad (2.6.2)$$

As with superiority, equivalence and non-inferiority trials discussed earlier in this chapter it is best to calculate the power using a non-central t-distribution the power, as outlined by Owen [1965], rewriting (2.6.1) to the following formula [Owen, 1965; Diletti, Hauschke, Steinijans, 1991; Julious, 2004a],

$$1-\beta = \text{probt}(-t_{1-\alpha,n-2}, n-2, \tau_2) - \text{probt}(t_{1-\alpha,n-2}, n-2, \tau_1), \quad (2.6.3)$$

where τ_1 and τ_2 are non centrality parameters defined as

$$\tau_1 = \frac{\sqrt{n}(\log(\mu_T/\mu_R)-\log(0.80))}{\sqrt{2\sigma_w^2}} \text{ and } \tau_2 = \frac{\sqrt{n}(\log(\mu_T/\mu_R)-\log(1.25))}{\sqrt{2\sigma_w^2}}.$$

An estimate of the sample size for μ_T/μ_R greater than unity can be obtained from the following

$$n = \frac{2\sigma_w^2(Z_{1-\beta}+Z_{1-\alpha})^2}{(\log(\mu_T/\mu_R)-\log(1.25))^2}, \quad (2.6.4)$$

which can be used to provide an initial value for the iterations. This equation provides reasonable approximations for $\mu_T/\mu_R \neq 1$, especially when the mean ratio becomes large relative to (0.80 to 1.25) as in such circumstances most of the Type II error comes from one test of two one sided tests. For quick calculations, for 90% power and a Type I error of 5%, the following can be used

$$n = \frac{17\sigma_w^2}{(\log(\mu_T \ \mu_R) - \log(1.25))^2}. \qquad (2.6.5)$$

Obviously for true ratios less than unity log(1.25) should be replaced by log(0.80).

2.7.1.3. Special Case of the Mean Ratio Equalling Unity

For the special case where the expected true mean difference is expected to be unity the sample size can be directly derived from the following formula

$$n = \frac{2\sigma_w^2(Z_{1-\beta/2} + Z_{1-\alpha})^2}{(\log(1.25))^2}. \qquad (2.6.6)$$

By replacing the Z-statistic with the t-statistics (2.6.6) can be rewritten to give the sample size as

$$n = \frac{2\sigma_w^2(Z_{1-\beta/2} + t_{1-\alpha,n-2})^2}{(\log(1.25))^2}. \qquad (2.6.7)$$

In turn this can be rewritten as

$$1-\beta = 2\Phi\left(\sqrt{\frac{(\log(1.25))^2 n}{2\sigma_w^2}} - t_{1-\alpha,n-2}\right) - 1.$$

Estimating the power from a non-central t-distribution, (2.6.3) can be rewritten to

$$1-\beta = 2\text{probt}(-t_{1-\alpha,n-2}, n-2, \tau) - 1, \qquad (2.6.8)$$

where τ is the non centrality parameters defined as

$$\tau = \frac{-\sqrt{n}(\log(1.25))}{\sqrt{2\sigma_w^2}}.$$

Equation (2.6.6) can be used to obtain initial estimates of the sample size to use in (2.6.8). For quick calculations for 90% power, 5% Type I error rate and 20% acceptance criteria one could use

$$n = 433\sigma_w^2. \qquad (2.6.9)$$

Table 2.23 gives sample size estimates using (2.6.3) for different CVs, mean ratios and acceptance criteria 10% (0.90 to 1.11), 15% (0.85 to 1.18), 20% (0.80 to 1.25) etc for a Type I error rate of 5% and 90% power. The simpler equations provide good estimates of the total sample size, underestimating the sample size by one or two, and hence good initial values for iteration.

Table 2-23. Total sample sizes (n) for bioequivalence cross-over study for different CVs, levels of bioequivalence and true mean ratios for 90% power and type I error of 5%

		Levels of Bio-equivalence				
CV (%)	Ratio	10%	15%	20%	25%	30%
20	0.80				163	40
	0.85			18	45	20
	0.90		207	50	22	13
	0.95	232	56	2	14	10
	1.00	78	34	1	12	9
	1.05	212	54	2	14	10
	1.10		151	43	20	12
	1.15			99	33	16
	1.20			405	62	24
25	0.80				163	40
	0.85			185	45	20
	0.90		2073	50	22	13
	1.00	232	56	25	14	10
	1.05	78	34	19	12	9
	1.05	212	54	24	14	10
	1.10		151	43	20	12
	1.15			99	33	16
	1.20			405	62	24
30	0.80				356	85
	0.85			403	96	41
	0.90		454	108	46	25
	0.95	507	121	52	29	18
	1.00	170	73	39	25	17
	1.05	463	116	51	28	18
	1.10		329	92	42	24
	1.15			214	69	33
	1.20			888	135	50
35	0.80				477	113
	0.85			540	128	54
	0.90		608	145	61	33
	0.95	679	162	69	38	24
	1.00	227	97	52	32	22
	1.05	620	155	67	37	24
	1.10		440	123	55	31
	1.15			287	92	44
	1.20			1190	180	67
40	0.80				612	144
	0.85			694	164	69
	0.90		780	185	78	42
	0.95	871	207	88	48	30
	1.00	291	124	66	41	27
	1.05	796	198	86	47	30
	1.10		565	157	71	39
	1.15			367	118	56
	1.20			1527	231	86

Note the "standard" bio-equivalence criteria are of a 20% difference on the log scale (0.80 to 1.25) but these need not always be used. For a drug with a narrow safety

window (with respect to dose) a narrower margin may be used; while for in-vivo assessment (drug-interactions, food effect) wider margins may be used – for example for food effect studies a margin of (0.70 to 1.43) for Cmax [FDA, 1997].

2.7.1.4. Replicate Designs

For compounds with high variability the standard AB/BA can require very large sample sizes, especially if the mean ratio is not expected to be unity. One type of design, which can partially overcome this problem, are replicate cross-over designs. By adding an extra arm to the study such that the sequences are ABB/BAA one can reduce the sample size by 25% compared to a standard AB/BA design; while an ABBA/BAAB design can reduce the sample size by 50% [Liu, 1995]. This option may not be practical for certain compounds, for example those with a long half-life, but it is a possible solution for compounds with high pharmacokinetic variability.

Another type of replicate design is a two period replicate design AA/AB/BA/BB – also known as Balaam's Design [Jones and Kenward, 2003]. This design allows for an intra-subject estimate of variability for a given compound without increasing the number of periods beyond two (more than two periods may not be practical). To consider the effect such a design has on the sample size one must consider the derivation of the total variance $\sigma^2 = \sigma_b^2 + \sigma_w^2$, where σ_w^2 is the within-subject component of variation and σ_b^2 is the between subject component of variation. Both these variance components can be estimated from previous cross-over trials with the test and reference compounds. Now suppose $\sigma_b^2 = k\sigma_w^2$ it can be shown, assuming an equal allocation to each sequence, that the sample size required for a two period replicate design can be derived by multiplying the sample size for standard AB/BA design as follows [Julious, 2004a]

$$n_{AA\ AB\ BA\ BB} = \left(\frac{2k+1}{k+1}\right) n_{AB\ BA}. \qquad (2.6.10)$$

The derivation number of this formula comes initially from imagining that the AB/BA and AA/BB sequences are from a cross-over trial and a parallel group trial respectively with n/4 subjects assigned to each sequence. For each sequences the following total variance can thus be derived for the "parallel group" sequences

$$\frac{4\sigma^2}{n} + \frac{4\sigma^2}{n}.$$

If these sequences were from a parallel group study one would effectively take the average of the two sessions to compared A and B and so from (2.8.5) given later in the chapter $\sigma^2 = \sigma_b^2 + \sigma_w^2/2$ and with $\sigma_b^2 = k\sigma_w^2$ this the variance becomes $4(2k+1)\sigma_w^2$ which equals w_1 say. Now for the "cross-over" AB/BA sequences, the total variance can be derived as

$$\frac{4\sigma_w^2}{n},$$

which equals w_2 say. Now to combine the cross-over and parallel sequences into one overall variance one could use the following formula borrowed from meta-analysis methodology [Whitehead and Whitehead, 1991, Julious, 2004a]

$$\left(\sum_{i=1}^{2}\frac{1}{w_i}\right)^{-1}.$$

Thus, the overall variance is [Julious, 2004a]

$$\left(\frac{n}{4\sigma_w^2}+\frac{n}{4\sigma_w^2(2k+1)}\right)^{-1}=\frac{2\sigma_w^2(2k+1)}{n(k+1)}.$$

From any of the sample size formulae given in this chapter it is evident that one increases the sample size directly proportional to any increase in the variance. If one is planning a simple AB/BA cross-over trial the overall variance would be $2\sigma_w^2/n$. Thus, the ratio of the variances is thus

$$\frac{2\sigma_w^2(2k+1)}{n(k+1)}\frac{n}{2\sigma_w^2}=\frac{(2k+1)}{(k+1)},$$

and so the increase in sample sizes for doing a replicate cross-over is

$$n_{AA\ AB\ BA\ BB}=\left(\frac{2k+1}{k+1}\right)n_{AB\ BA},$$

and (2.6.10).

To verify this result 10,000 simulations for a fixed sample size of 48 and for various k were undertaken. Each simulation simulated AB/BA and AB/BA/AA/BB cross-over. The analysis for each simulation was done with all subjects entered into PROC MIXED with subject entered as random. Table 2.24 gives the results.

Table 2-24. Multiplication factors for different values of k for a two period replicate cross-over design

k	$\frac{2k+1}{k+1}$	Simulation
2	1.67	1.65
4	1.80	1.78
6	1.86	1.85
8	1.89	1.88
10	1.91	1.90

What is evident both from the table above and (2.6.10) that for a two period replicate design will always require more subjects than a standard AB/BA requiring the same sample size only for k=0. However, no matter how larger k becomes it will only require twice as many subjects at most. This is because as k becomes large virtually all the information, in the comparison of the mean ratio, comes from the AB/BA sequences and with twice as many subjects there will be as many people in these sequences as in a standard AB/BA design.

2.7.1.5. Worked Example

A bioequivalence trial to compare a test with reference formulation needs to be designed. The standard bioequivalence criteria will be used to demonstrate that the average drug exposure on the test is bioequivalent to the reference i.e. 0.80 to 1.25. The within-subject standard deviation is expected to be 0.25 ($=\sigma_w$) and the mean ratio is expected to be unity ($\mu_T/\mu_R = 1$). The standard deviation of 0.25 equates to a within subject CV of 25%. The study design will be an AB/BA two period crossover. From Table 2.23 it can be seen that one would need a minimum evaluable sample size of 28 subjects. Practically this would equate to at least 28 subjects in total or 14 subjects on each sequence (AB and BA) maybe with approximately 20% more subjects added to the sample size (i.e. 17 subjects per sequence) to account for drop outs.

If the test formulation is expected, on average, to have exposures 5% greater than the reference ($\mu_T/\mu_R = 1.05$) then the total sample size would increase to 36 subjects (or 18 per sequence).

Suppose though instead of an AB/BA design replicate ABB/BAA or ABBA/BAAB designs were being considered for the case where exposures were expected to be 5% greater on test compared to reference. If one adapted a 4 period replicate design then would would multiply the total sample size calculated earlier by 0.50 to get 36x0.5=18 subjects in total required. If one adopted a 3 period replicate design then the total sample size calculated earlier should be multiplied by 0.75 to get 36x0.75=27 subjects in total.

2.7.1.6. Sensitivity Analysis About the Variance Used in the Sample Size Calculations

As with other types of trial described earlier in this chapter (2.2.9) can be used to estimate plausibly large value for the population variance. For bioequivalence trials (as with equivalence trials), one has the further factor to investigate of the sensitivity of calculations about the true mean ratio. If one has assumed a mean ratio of unity then it will affect the power of the study if this assumption is false.

Table 2-25. Sample sizes for a bioequivalence study for different mean ratios assuming 90% power and 5% type I error rate along with the powers corresponding to the 95th percentile of the variance for different degrees of freedom and different true mean ratios

		Sample	Degrees of Freedom											
CV(%)	Ratio	Size	Ratio	5	10	15	20	25	30	40	50	75	100	∞
20	1.00	19	1.00	0.00	0.35	0.50	0.58	0.63	0.67	0.71	0.74	0.78	0.80	0.91
			1.05	0.00	0.31	0.45	0.52	0.56	0.59	0.63	0.66	0.70	0.72	0.82
			1.10	0.00	0.24	0.33	0.37	0.40	0.42	0.45	0.47	0.50	0.51	0.60
			1.15	0.00	0.15	0.20	0.22	0.24	0.25	0.26	0.27	0.28	0.29	0.34
	1.05	24	1.05	0.12	0.48	0.60	0.66	0.70	0.73	0.76	0.78	0.81	0.82	0.90
			1.10	0.10	0.35	0.43	0.47	0.50	0.52	0.54	0.56	0.59	0.60	0.70
			1.15	0.07	0.21	0.25	0.27	0.29	0.30	0.31	0.32	0.34	0.35	0.41
	1.10	43	1.10	0.38	0.58	0.66	0.70	0.73	0.75	0.77	0.79	0.82	0.83	0.90
			1.15	0.22	0.33	0.38	0.41	0.43	0.45	0.47	0.48	0.51	0.52	0.61
	1.15	99	1.15	0.41	0.58	0.66	0.70	0.73	0.75	0.77	0.79	0.82	0.83	0.90
25	1.00	28	1.00	0.00	0.33	0.49	0.57	0.62	0.66	0.70	0.73	0.77	0.80	0.90
			1.05	0.00	0.30	0.44	0.51	0.55	0.58	0.62	0.65	0.69	0.71	0.82
			1.10	0.00	0.23	0.32	0.37	0.40	0.42	0.44	0.46	0.49	0.50	0.60
			1.15	0.00	0.15	0.20	0.22	0.23	0.24	0.26	0.27	0.28	0.29	0.34
	1.05	36	1.05	0.12	0.47	0.60	0.66	0.70	0.72	0.76	0.78	0.81	0.82	0.90
			1.10	0.09	0.34	0.43	0.47	0.50	0.51	0.54	0.56	0.59	0.60	0.70
			1.15	0.07	0.21	0.25	0.27	0.28	0.29	0.31	0.32	0.33	0.34	0.41
	1.10	65	1.10	0.37	0.57	0.65	0.70	0.72	0.74	0.77	0.79	0.81	0.83	0.90
			1.15	0.22	0.33	0.38	0.41	0.43	0.44	0.46	0.48	0.50	0.52	0.60
	1.15	151	1.15	0.40	0.58	0.65	0.70	0.72	0.74	0.77	0.79	0.81	0.83	0.90
30	1.00	39	1.00	0.00	0.33	0.48	0.57	0.62	0.65	0.70	0.73	0.77	0.79	0.90
			1.05	0.00	0.30	0.43	0.50	0.55	0.58	0.62	0.65	0.69	0.71	0.81
			1.10	0.00	0.23	0.32	0.37	0.39	0.41	0.44	0.46	0.49	0.50	0.59
			1.15	0.00	0.15	0.20	0.22	0.23	0.24	0.26	0.27	0.28	0.29	0.34
	1.05	51	1.05	0.12	0.48	0.60	0.66	0.70	0.73	0.76	0.78	0.81	0.83	0.90
			1.10	0.10	0.35	0.43	0.47	0.50	0.52	0.54	0.56	0.59	0.61	0.70
			1.15	0.07	0.21	0.25	0.27	0.29	0.30	0.31	0.32	0.34	0.35	0.41
	1.10	92	1.10	0.37	0.58	0.65	0.70	0.72	0.74	0.77	0.79	0.81	0.83	0.90
			1.15	0.22	0.33	0.38	0.41	0.43	0.44	0.46	0.48	0.50	0.52	0.61
	1.15	214	1.15	0.40	0.58	0.65	0.70	0.72	0.74	0.77	0.79	0.81	0.83	0.90

Table 2.25 gives the sample size per group required for different CVs for a cross-over bioequivalence trial (assuming the standard 20% bioequivalence criteria are being used) along with the sensitivity of these sample sizes to the 95th percentile of the variance (for different degrees of freedom) and different mean ratios. From Table 2.25 it seems:

- That to ensure that 50% and 80% power with the 95th percentile about the variance one would require 15 and 100 degrees of freedom.

- If one's variance estimate was close to the true variance then even if the true mean ratio 1.05 one would still have 80% in the study if designed under the assumption of unity.

- With moderate degrees of freedom a study is reasonably robust to deviations in the assumptions either about the true mean or the variance but less so to deviations in both simultaneously.

2.7.1.7. Worked Example

Revisiting the example given earlier. Suppose that the CV used in the sample size calculations was estimated with just 15 degrees of freedom. From Table 2.25 it seems that for a high plausible value for the CV the study has 49% power. If the true mean ratio is really 1.05 as opposed to 1.00 then the study would have 82% power.

2.7.1.8. Calculations Taking of Account the Imprecision of the Variance Used in the Sample Size Calculations

2.7.1.9. General Case

Extending the arguments for equivalence trials given earlier in this chapter the sample size for a bioequivalence study, taking into account the degrees of freedom about the sample variance study, can be derived from

$$1-\beta = \text{probt}\left(-\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(1.25))^2 n}{2s_w^2}}, df, -t_{1-\alpha,n-2}\right) - \text{probt}\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(0.80))^2 n}{2s_w^2}}, df, t_{1-\alpha,n-2}\right), \quad (2.6.11)$$

where s_w^2 is a sample estimate of the within subject population variance. Replacing the t-statistic with the z-statistic and (2.6.11) becomes

$$1-\beta = \text{probt}\left(-\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(1.25))^2 n}{2s_w^2}}, df, -Z_{1-\alpha}\right) + \text{probt}\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(0.80))^2 n}{2s_w^2}}, df, Z_{1-\alpha}\right) - 1. \quad (2.6.12)$$

A direct estimate of the sample size can be obtained as an initial estimate for (2.6.11) for the expected true mean ratio becomes large, $\mu_T/\mu_R \geq 1.05$. Hence, the following quick formula can be used to obtain direct initial estimates of the sample size for the general case of $\mu_T/\mu_R \neq 1$

$$n = \frac{2s_w^2[tinv(1-\beta, df, Z_{1-\alpha})]^2}{[\log(1.25)-\log(\mu_T/\mu_R)]^2}. \quad (2.6.13)$$

Table 2-26. Sample sizes for bioequivalence cross-over studies for various CVs and degrees of freedom using (2.6.11), for 90% power and 5% type I error rate assuming 20% (0.80 to 1.25) bioequivalence limits. The row with "infinite" degrees of freedom is from (2.6.2)

		Coefficients of Variation				
Ratio	m	20	25	30		40
1.00	5	35	54	76	101	129
	10	25	38	54	71	91
	15	23	34	48	64	82
	20	22	33	46	61	77
	25	21	32	44	59	75
	30	21	31	43	57	73
	40	20	30	42	56	71
	50	20	30	41	55	70
	75	19	29	41	54	69
	100	19	29	40	53	68
	∞	19	28	39	52	66
1.05	5	43	65	91	122	156
	10	31	47	67	89	114
	15	28	43	60	80	103
	20	27	41	58	77	98
	25	26	40	56	75	95
	30	26	39	55	73	94
	40	25	38	54	72	92
	50	25	38	53	71	90
	75	25	37	52	69	89
	100	24	37	52	69	88
	∞	24	36	50	67	86
1.10	5	71	108	153	205	263
	10	54	83	117	157	201
	15	50	76	108	144	184
	20	48	73	104	138	177
	25	47	72	101	135	173
	30	46	70	99	133	170
	40	45	69	97	130	167
	50	45	68	96	129	165
	75	44	67	95	127	162
	100	44	67	94	126	161
	∞	43	65	92	123	157

2.7.1.10. Special Case of the Mean Ratio Equalling Unity

For the special case of $\mu_I = \mu_R$ (2.6.11) can be rewritten as

$$1-\beta = 2\text{probt}\left(\frac{\log(0.80)\sqrt{n}}{\sqrt{2s_w^2}}, df, -t_{1-\alpha,df}\right) - 1, \tag{2.6.14}$$

which when replacing the t statistic with the Z statistics becomes

$$1-\beta = 2\text{probt}\left(\frac{\log(0.80)\sqrt{n}}{\sqrt{2s_w^2}}, df, -Z_{1-\alpha}\right) - 1. \tag{2.6.15}$$

Hence, a direct estimate of the sample size can be obtained from

$$n = \frac{2s_w^2[tinv(1-\beta/2, df, Z_{1-\alpha})]^2}{(\log(1.25))^2}. \qquad (2.6.16)$$

Table 2.26 gives estimates of the sample size for bioequivalence cross-over studies using the standard 20% bioequivalence criteria.

Table 2-27. Multiplication factors for different levels of one sided significance, type II error and degrees of freedom

		Significance Level (α)			
m	β	0.010	0.025	0.050	0.100
5	0.05	2.649	2.509	2.385	2.238
	0.10	2.167	2.068	1.980	1.875
	0.15	1.929	1.850	1.780	1.696
	0.20	1.776	1.711	1.652	1.581
	0.50	1.367	1.337	1.311	1.278
10	0.05	1.611	1.562	1.520	1.470
	0.10	1.463	1.425	1.392	1.353
	0.15	1.382	1.351	1.323	1.290
	0.20	1.328	1.301	1.276	1.248
	0.50	1.166	1.153	1.141	1.127
25	0.05	1.208	1.192	1.178	1.162
	0.10	1.163	1.150	1.139	1.125
	0.15	1.137	1.126	1.116	1.105
	0.20	1.119	1.109	1.101	1.091
	0.50	1.062	1.058	1.053	1.058
50	0.05	1.099	1.091	1.085	1.077
	0.10	1.078	1.072	1.067	1.060
	0.15	1.066	1.061	1.056	1.051
	0.20	1.058	1.053	1.049	1.044
	0.50	1.031	1.028	1.026	1.024
75	0.05	1.065	1.060	1.056	1.051
	0.10	1.052	1.047	1.044	1.040
	0.15	1.044	1.040	1.037	1.033
	0.20	1.038	1.035	1.032	1.029
	0.50	1.020	1.019	1.017	1.016
100	0.05	1.048	1.044	1.041	1.038
	0.10	1.038	1.035	1.033	1.030
	0.15	1.033	1.030	1.028	1.025
	0.20	1.029	1.026	1.024	1.022
	0.50	1.015	1.014	1.013	1.012

Table 2.27 gives the multiplication factors, compared to assuming one has the population variance, for various degrees of freedom and Type I and II errors assuming a mean ratio of unity. This table is in fact the same as Table 2.12 given for equivalence trials - although for bioequivalence trials one would choose different Type I errors. Table 2.12 (and Table 2.27) can be used regardless of which original sample size formula was used.

2.7.1.11. Worked Example

Revisiting the example given earlier with just 15 degrees of freedom of the sample variance used in calculations one would require 32 subjects in the trial as opposed to 28 calculated previously. This is an increase in the sample required of 15%.

2.7.1.12. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations

Similarly to equivalence trials with an estimate of the mean ratio ($\bar{x}_T / \bar{x}_R$) from a previous study and an estimate of the standard error around this mean ratio on the log scale ($se\log(\bar{x}_T / \bar{x}_R)$) an estimate of the sample size can be obtained. For the special case $\bar{x}_T / \bar{x}_R = 1$ the power for a given sample size can be estimated (through numerical integration) from

$$1-\beta = \sum_{p=0.001}^{0.998} 0.001x \left[\begin{array}{l} 2\text{Prob}t\left(\dfrac{\sqrt{n}\left(Z_p se\log(\bar{x}_T / \bar{x}_R) - \log(1.25)\right)}{\sqrt{2s_w^2}}, m, -t_{1-\alpha,n-2} \right) \\ +2\text{Prob}t\left(\dfrac{\sqrt{n}\left(Z_p se\log(\bar{x}_T / \bar{x}_R) - \log(1.25)\right)}{\sqrt{2s_w^2}}, m, -t_{1-\alpha,n-2} \right) - 2 \end{array} \right] / 2, \qquad (2.6.17)$$

where the sample size required is the minimum value which gives the required sample size. As with equivalence trials this result, however, is practically unappealing.

2.7.1.13. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations – A Bayesian Approach

In this section it will be demonstrated how simple Bayesian methods could be employed to estimate sample sizes allowing for imprecision in the mean and variance.

2.7.1.14. Prior Response

Prior values for the mean difference and population within subject standard deviation on the log scale are defined as d_0 (i.e. for $\log(\bar{x}_T) - \log(\bar{x}_T)$) and s_{w0} respectively.

2.7.1.15. Anticipated Response

The anticipated mean difference and population within subject standard deviation on the log scale are defined as d_1 and s_{w1} respectively. These values are taken as objective values observed in a previously conducted bioequivalence trial.

2.7.1.16. Posterior Response

With these anticipated and prior responses the posterior distribution can be calculated, with the posterior estimate of the population variance around the mean, s^2_{wn}, is defined as

$$s^2_{wn} = \left(\frac{n}{2s^2_{w1}} + \frac{1}{s_{w0}} \right)^{-1}, \qquad (2.6.18)$$

and the posterior estimate of the mean difference, d_n, is defined as

$$d_n = s^2_{wn} \left(\frac{d_{w0}}{s^2_{w0}} + \frac{d_1 n}{2s^2_1} \right). \qquad (2.6.19)$$

From these posterior values a density distribution for $prob(\theta > d_i | d_1)$ can be defined such that a probability of observing d_i, or greater, for a given d_n would be

$$prob(\theta > d_i | d_1) = \Phi\left(\frac{d_n - d_i}{s_{wn}} \right), \qquad (2.6.20)$$

or alternatively for $\Pr ob(\theta < d_i | d_1)$

$$prob(\theta > d_i | d_1) = \Phi\left(\frac{d_i - d_n}{s_{wn}} \right).$$

From this result a log-mean difference d_i for given percentile p can be estimated and put into the following result for the general case $\log(\mu_T / \mu_T) \neq 0$, to estimate the power for a given sample size

$$1 - \beta = \sum_{p=0.001}^{0.999} 0.001 \left\{ \left[\text{Probt}\left(\frac{\sqrt{n}(d_p - d)}{\sqrt{2s^2_w}}, m, -t_{1-\alpha,2} \right) + \text{Probt}\left(\frac{\sqrt{n}(d_p + d)}{\sqrt{2s^2}}, m, t_{1-\alpha,2} \right) \right] + \left[\text{Probt}\left(\frac{\sqrt{n}(d_{p=0.001} - d)}{\sqrt{2s^2_w}}, m, -t_{1-\alpha,2} \right) + \text{Probt}\left(\frac{\sqrt{n}(d_p + d)}{\sqrt{2s^2}}, m, t_{1-\alpha,2} \right) \right] \right\} / 2, \qquad (2.6.21)$$

where the sample size is estimated through numerical integration and iteration.

2.7.2. Parallel Group Studies

Although cross-over trials are the 'norm' for the assessment of bioequivalence sometimes, particularly with very long half life compounds, these designs are not practical. This section briefly describes the methodology for sample size calculation for parallel group bioequivalence trials.

2.7.2.1. Sample Size Estimated Assuming the Population Variance to be Known

2.7.2.2. General Case

The power for a bioequivalence trial with acceptance limits of (0.8, 1.25) for given values of the any true ratio is given by

$$1-\beta=\Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(1.25))^2 rn_T}{(r+1)\sigma^2}}-Z_{1-\alpha}\right)+\Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(0.80))^2 rn_T}{(r+1)\sigma^2}}-Z_{1-\alpha}\right)-1, \qquad (2.6.22)$$

where σ is the between-subject variability on the log scale, r is the allocation ratio and n_T is the sample size in the test group. Replacing the Z-statistic with a t-statistics and (2.6.22) can be rewritten as

$$1-\beta=\Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(1.25))^2 rn_T}{(r+1)\sigma^2}}-t_{1-\alpha,n_T(r+1)-2}\right)+\Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(0.80))^2 rn_T}{(r+1)\sigma^2}}-t_{1-\alpha,n_T(r+1)-2}\right)-1, \qquad (2.6.23)$$

and under the assumption of a non-central t-distribution the power is estimated from

$$1-\beta=\text{Probt}\left(-t_{1-\alpha,n_T(r+1)-2},n_T(r+1)-2,\tau_2\right)-\text{Probt}\left(t_{1-\alpha,n_T(r+1)-2},n_T(r+1)-2,\tau_1\right), \qquad (2.6.24)$$

where τ_1 and τ_2 are non centrality parameters defined as

$$\tau_1=\frac{\sqrt{rn_T}\left(\log(\mu_T/\mu_R)-\log(0.80)\right)}{\sqrt{(r+1)\sigma^2}} \text{ and } \tau_2=\frac{\sqrt{rn_T}\left(\log(\mu_T/\mu_R)-\log(1.25)\right)}{\sqrt{(r+1)\sigma^2}}.$$

As with a cross-over trial a direct estimate of the sample size for a mean ratio greater than unity can be obtained from the following

$$n_T=\frac{(r+1)\sigma^2\left(Z_{1-\beta}+Z_{1-\alpha}\right)^2}{r\left(\log(\mu_T/\mu_R)-\log(1.25)\right)^2}, \qquad (2.6.25)$$

and for quick calculations one could use

$$n_T=\frac{17(r+1)\sigma^2}{r\left(\log(\mu_T/\mu_R)-\log(1.25)\right)^2}. \qquad (2.6.26)$$

If the mean ratio is expected to be less than unity then replace log(1.25) with log(0.80) in (2.6.25) and (2.6.26).

Table 2-28. Sample sizes for one arm of a bioequivalence parallel group study for different CVs, levels of bioequivalence and true mean ratios for 90% power and a type I error rate of 5%

		Levels of Bio-equivalence				
CV (%)	Ratio	10%	15%	20%	25%	30%
30	0.80				356	84
	0.85			40	95	40
	0.90		453	10	46	25
	0.95	506	121	5	28	18
	1.00	169	72	3	24	16
	1.05	462	115	5	28	17
	1.10		328	92	41	23
	1.15			21	69	33
	1.20			887	134	50
35	0.80				476	112
	0.85			540	128	54
	0.90		607	144	61	33
	0.95	678	161	69	37	23
	1.00	226	96	51	31	21
	1.05	620	154	67	37	23
	1.10		439	122	55	30
	1.15			286	92	43
	1.20			1189	179	66
40	0.80				611	144
	0.85			693	163	69
	0.90		779	184	78	41
	0.95	871	207	88	48	30
	1.00	291	123	66	40	26
	1.05	796	198	85	47	29
	1.10		564	157	70	38
	1.15			367	117	55
	1.20			1527	230	85
45	0.80				759	178
	0.85			861	203	85
	0.90		968	229	96	51
	0.95	1082	257	109	59	36
	1.00	361	152	81	49	33
	1.05	988	245	106	58	36
	1.10		700	194	87	47
	1.15			455	146	68
	1.20			1896	286	105
50	0.80				919	216
	0.85			1041	245	103
	0.90		1171	277	116	62
	0.95	1309	310	131	71	44
	1.00	436	184	98	60	39
	1.05	1195	297	128	70	43
	1.10		847	235	104	57
	1.15			551	176	82
	1.20			2295	345	127

2.7.2.3. Special Case of the Ratio Equalling Unity

When the mean ratio is expected to be unity the sample size can be derived directly from

$$n_T = \frac{(r+1)\sigma^2\left(Z_{1-\beta/2} + Z_{1-\alpha}\right)^2}{r\left(\log(1.25)\right)^2}. \qquad (2.6.27)$$

Replacing the Z-statistic with the t-statistic (2.6.27) can be rewritten as

$$n_T = \frac{(r+1)\sigma^2\left(Z_{1-\beta/2} + t_{1-\alpha,n_T(r+1)-2}\right)^2}{r\left(\log(1.25)\right)^2}. \qquad (2.6.28)$$

Equation (2.6.28) can in turn can be rewritten as

$$1-\beta = 2\Phi\left(\sqrt{\frac{\left(\log(1.25)\right)^2 r n_T}{(r+1)\sigma^2}} - t_{1-\alpha,n_T(r+1)-2}\right) - 1,$$

and under the assumption of a non-central t-distribution the power can be derived from

$$1-\beta = 2\text{probt}\left(-t_{1-\alpha,n_T(r+1)-2}, n_T(r+1)-2, \tau\right) - 1, \qquad (2.6.29)$$

where τ is the non centrality parameters defined as

$$\tau = \frac{-\sqrt{n_T}\, r\left(\log(1.25)\right)}{\sqrt{(r+1)\sigma^2}}.$$

Equation (2.6.27) can be used for initial estimates of the sample size to use in (2.6.29). For quick calculations of the sample size for 90% power, 5% Type I error rate and 20% one could use

$$10.75(r+1)\sigma^2/r. \qquad (2.6.30)$$

Table 2.28 gives sample size estimates using (2.6.24) for different CVs, mean ratios and acceptance criteria 10% (0.90 to 1.11), 15% (0.85 to 1.18), 20% (0.80 to 1.25) etc for a Type I error rate of 5%, 90% power and an allocation ratio of one. As with cross-over trials the simpler equations provide good estimates for initial calculations.

2.7.2.4. Sensitivity Analysis About the Variance Used in the Sample Size Calculations

As with the cross-over trials described earlier an example table such as Table 2.29 can be used to investigate the sensitivity of a parallel group bioequivalence trial. The inference from this table is the same as that for cross-over trials.

Table 2-29. Sample sizes for a parallel group bioequivalence study with different mean ratios assuming 90% power and 5% type I error rate along with the powers corresponding to the 95th percentile of the variance for different degrees of freedom and different true mean ratios

		Sample		Degrees of Freedom										
CV (%)	Ratio	Size	Ratio	5	10	15	20	25	30	40	50	75	100	∞
50	1.00	98	1.00	0.00	0.33	0.48	0.56	0.62	0.65	0.70	0.73	0.77	0.79	0.90
			1.05	0.00	0.30	0.43	0.50	0.55	0.58	0.62	0.65	0.68	0.71	0.81
			1.10	0.00	0.23	0.32	0.36	0.39	0.41	0.44	0.46	0.49	0.50	0.59
			1.15	0.00	0.15	0.20	0.22	0.23	0.24	0.26	0.27	0.28	0.29	0.34
	1.05	128	1.05	0.11	0.47	0.60	0.66	0.70	0.72	0.76	0.78	0.81	0.82	0.90
			1.10	0.09	0.34	0.43	0.47	0.50	0.52	0.54	0.56	0.59	0.60	0.70
			1.15	0.07	0.21	0.25	0.27	0.28	0.29	0.31	0.32	0.33	0.34	0.41
	1.10	235	1.10	0.37	0.58	0.65	0.70	0.72	0.74	0.77	0.79	0.81	0.83	0.90
			1.15	0.22	0.33	0.38	0.41	0.43	0.44	0.46	0.48	0.50	0.52	0.60
	1.15	551	1.15	0.40	0.58	0.65	0.70	0.72	0.74	0.77	0.79	0.81	0.83	0.90

2.7.2.5. Calculations Taking of Account the Imprecision of the Variance Used in the Sample Size Calculations

2.7.2.6. General Case

For a parallel group bioequivalence study the sample size can be derived from

$$1-\beta = \text{probt}\left(-\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(1.25))^2 rn_T}{(r+1)s^2}}, df, -t_{1-\alpha,(r+1)-2}\right) + \text{probt}\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(0.80))^2 rn_T}{(r+1)s^2}}, df, t_{1-\alpha,(r+1)-2}\right), \quad (2.6.31)$$

where s^2 is a sample estimate of the population variance. Replacing the t-statistic with the z-statistic and (2.6.31) becomes

$$1-\beta = \text{probt}\left(-\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(1.25))^2 rn_T}{(r+1)s^2}}, df, -Z_{1-\alpha}\right) + \text{probt}\left(\sqrt{\frac{(\log(\mu_T/\mu_R)-\log(0.80))^2 rn_T}{(r+1)s^2}}, df, Z_{1-\alpha}\right). \quad (2.6.32)$$

For $\mu_T/\mu_R \geq 1.05$ a direct estimate of the sample size can to start iterations can be obtained from

$$n = \frac{(r+1)s^2[tinv(1-\beta, df, Z_{1-\alpha})]^2}{r[\log(1.25)-\log(\mu_T/\mu_R)]^2}. \quad (2.6.33)$$

2.7.2.7. Special Case of the Mean Ratio Equaling Unity

For the special case of $\mu_T = \mu_R$ (2.6.31) can be rewritten as

$$1-\beta = 2\text{probt}\left(-\sqrt{\frac{(\log\log(1.25))^2 rn_T}{(r+1)s^2}}, df, -t_{1-\alpha,(r+1)-2}\right) - 1. \quad (2.6.34)$$

which when replacing the t-statistic with the Z-statistics becomes

$$1-\beta = 2\text{probt}\left(-\sqrt{\frac{(\log\log(1.25))^2 rn_1}{(r+1)s^2}}, df, -Z_{1-\alpha}\right) - 1. \qquad (2.6.35)$$

Hence, a direct estimate of the sample size can be obtained from

$$n = \frac{2s_w^2[tinv(1-\beta/2, df, Z_{1-\alpha})]^2}{(\log(1.25))^2}, \qquad (2.6.36)$$

Table 2.30 gives estimates of the sample size for bioequivalence parallel group studies using the standard 20% bioequivalence criteria.

Table 2-30. Sample sizes for bioequivalence parallel group studies for various CVs and degrees of freedom using (2.6.25), for 90% power and 5% type I error rate assuming 20% (0.80 to 1.25) bioequivalence limits. The row with "infinite" degrees of freedom is from (2.6.18)

Ratio	Degrees of Freedom	Coefficient of Variation 30	35	40	45	50
1.00	5	101	129	160	193	229
	10	71	91	113	136	161
	15	63	81	101	122	144
	20	60	77	95	115	136
	25	58	74	92	111	132
	30	57	73	90	109	129
	40	55	71	88	106	126
	50	55	70	86	104	124
	75	53	68	85	102	121
	100	53	68	84	101	120
	∞	51	65	81	98	116
1.05	5	121	156	193	233	276
	10	88	113	140	169	200
	15	80	102	127	153	181
	20	76	98	121	146	173
	25	74	95	118	142	168
	30	73	93	115	140	165
	40	71	91	113	136	161
	50	70	90	111	135	159
	75	69	88	109	132	156
	100	68	87	108	131	155
	∞	67	85	106	128	151
1.10	5	20	262	326	394	466
	10	15	200	248	300	355
	15	14	184	228	276	327
	20	13	177	219	265	314
	25	13	172	214	259	306
	30	13	170	210	254	301
	40	13	166	206	249	295
	50	12	164	204	246	292
	75	12	162	201	242	287
	100	12	160	199	241	285
	∞	12	157	194	235	278

Table 2.27 (given earlier for cross-over trials) can be used for multiplication factors, as these depend only on degrees of freedom and Type I and II errors.

2.7.2.8. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations

Similarly to cross-over trials with an estimate of the mean ratio and an estimate of the standard error around this mean ratio on the log scale ($se\log(\bar{x}_T/\bar{x}_R)$) an estimate of the sample size can be obtained. For the special case $\bar{x}_T/\bar{x}_R=1$ the power for a given sample size is

$$1-\beta=\sum_{p=0.001}^{0.998}0.001x\left[\begin{array}{l}2\text{Probt}\left(\dfrac{\sqrt{rn_i}\left(Z_p se\log(\bar{x}_T/\bar{x}_R)-\log(1.25)\right)}{\sqrt{(r+1)s}},m,t_{1-\alpha,n_i(r+1)-2}\right)\\+2\text{Probt}\left(\dfrac{\sqrt{rn_i}\left(Z_{p+0.001}se\log(\bar{x}_T/\bar{x}_R)-\log(1.25)\right)}{\sqrt{(r+1)s}},m,t_{1-\alpha,n_i(r+1)-2}\right)-2\end{array}\right]/2. \qquad (2.6.37)$$

2.7.2.9. Calculations Taking Account of the Imprecision of the Mean and Variance Used in the Sample Size Calculations – A Bayesian Approach

With the similar derivation as for cross-over bioequivalence studies the following result for the general case of $\log(\bar{x}_T/\bar{x}_R)\neq 0$ can be used to calculate the power for a given sample size.

$$1-\beta=\sum_{p=0.001}^{0.998}0.001x\left(\left[\text{Probt}\left(\frac{\sqrt{rn_i}(d+d_p)}{\sqrt{(r+1)s^2}},m,t_{1-\alpha,n_i(r+1)-2}\right)+\text{Probt}\left(\frac{\sqrt{rn_i}(d+d_p)}{\sqrt{(r+1)s^2}},m,t_{1-\alpha,n_i(r+1)-2}\right)-1\right]+\left[\text{Probt}\left(\frac{\sqrt{rn_i}(d-d_{p+0.001})}{\sqrt{(r+1)s^2}},m,t_{1-\alpha,n_i(r+1)-2}\right)+\text{Probt}\left(\frac{\sqrt{rn_i}(d-d_{p+0.001})}{\sqrt{(r+1)s^2}},m,t_{1-\alpha,n_i(r+1)-2}\right)-1\right]\right)/2, \qquad (2.3.38)$$

with the sample size estimated through iteration.

2.8. Estimation to a Given Precision

2.8.1. Parallel Group Trials

2.8.1.1. Sample Size Estimated Assuming the Population Variance to be Known

As discussed in Chapter 1 a $(1-\alpha)100\%$ confidence interval for $f(\mu)$ has half-width

$$w=Z_{\alpha\cdot 2}\sqrt{Var\,(S)}, \qquad (2.7.1)$$

and so defining Var(S) as per equation

$$Var\,(S)=\frac{\sigma^2}{n_A}+\frac{\sigma^2}{n_B}=\frac{r+1}{r}\cdot\frac{\sigma^2}{n_A},$$

one can solve (2.7.1) to give [Brush, 1988; Day, 1988; Desu and Raghavarao, 1990; Julious, 2004a; Julious and Patterson, 2004]

$$n_A = \frac{(r+1)Z^2_{1-\alpha/2}\sigma^2}{rw^2}, \qquad (2.7.2)$$

If the population variance is to be assumed unknown in the statistical analysis (2.7.2) can be rewritten as [Julious, 2004a; Julious and Patterson, 2004]

$$n_A \geq \frac{(r+1)t^2_{1-\alpha/2,n_A(r+1)-2}\sigma^2}{rw^2}. \qquad (2.7.3)$$

Equation (2.7.3) can be solved iteratively to find a value of n_A where the left hand side of the equation is greater than the right. An alternative equation to solve for n_A would be

$$0.5 \geq \Phi\left(\sqrt{\frac{rn_Aw^2}{(r+1)\sigma^2}} - t_{1-\alpha/2,n_A(r+1)-2}\right). \qquad (2.7.4)$$

Equation (2.7.4) holds as if one was to rewrite (2.7.4) in terms of n one would first have

$$Z_{0.5} = 0 \geq \sqrt{\frac{rn_Aw^2}{(r+1)\sigma^2}} - t_{1-\alpha/2,n_A(r+1)-2},$$

and hence (2.7.3). Equation (2.7.4) is in fact the same as (2.2.5), given in the section in superiority trials, but with the Type II error set at 0.5 – although obviously as precision trials are not powered they cannot have any Type II error. The practical application of this result is given later in the section on sample sizes where the population variance is assumed unknown for calculations.

To allow for the Normal approximation (2.7.2) can have a correction factor added to assist in initial calculations [Guenther, 1981; Julious and Patterson, 2004]

$$n_A = \frac{(r+1)\sigma^2Z^2_{1-\alpha/2}}{rw^2} + \frac{Z^2_{1-\alpha/2}}{4}, \qquad (2.7.5)$$

and the following quick formula can be used assuming one wishes to have a 95% confidence interval for the precision estimates

$$n_A = \frac{4\sigma^2}{w^2}\frac{(r+1)}{r}, \qquad (2.7.6)$$

or for r=1

$$n_A = \frac{8\sigma^2}{w^2}.$$

Table 2.31 gives sample sizes using (2.7.3) for various standardised widths ($\delta = w/\sigma$). The simpler equations slightly underestimate the sample size.

Table 2-31. Sample sizes for one group, n_A (n_B=rn_A) in a parallel group study for different standardised widths and allocation ratios with 95% confidence intervals for the precision estimates

	Allocation ratios			
δ	1	2	3	4
0.10	770	578	513	481
0.20	194	145	129	121
0.30	87	65	58	54
0.40	50	37	33	31
0.50	32	24	22	20
0.60	23	17	15	14
0.70	17	13	12	11
0.80	14	10	9	9
0.90	11	8	7	7
1.00	9	7	6	6

2.8.1.2. Worked Example

An investigator wishes to design a pilot hypertension trial with equal allocation between groups where the objective is to estimate any possible effect on blood pressure of new treatment compared to control with precision around the point estimate of ±2mmHg (w). The expected standard deviation in the population in which the trial is to be undertaken is 20mmHg (σ). Thus, the standardised width equates to $\delta = d/\sigma = 2/20 = 0.10$. Table 2.31 gives a sample size of 770 patients in each arm of the trial.

2.8.1.3. Sensitivity Analysis About the Variance Used in the Sample Size Calculations

When undertaking sensitivity analysis in a study designed to estimate effects to a given precision it is not the power one investigates against a high plausible value for the variance but the precision itself.

Pecision based studies are usually undertaken early in the development of a compound, when, by definition, there is little variability information available. Hence, a sensitivity analysis of the study may be quite important.

Table 2.32 gives sample sizes and precision estimates for high plausible values of the variance for different standardised widths. From this table it seems that with few degrees of freedom (about the variance used in the calculations) the precision could be half that used for the sample size calculations for a high plausible value for the variance.

Table 2-32. Sample sizes for one group, n_A (n_B=rn_A) in a parallel group study for different standardised widths along with the precision for high plausible values for the variance

	Sample	Degrees of Freedom				
δ	Size	5	10	25	50	100
0.10	770	0.21	0.16	0.13	0.12	0.11
0.20	194	0.42	0.32	0.26	0.24	0.23
0.30	87	0.63	0.48	0.39	0.36	0.34
0.40	50	0.84	0.64	0.52	0.48	0.45
0.50	32	1.04	0.80	0.65	0.60	0.57
0.60	23	1.25	0.96	0.78	0.72	0.68
0.70	17	1.46	1.12	0.92	0.84	0.79
0.80	14	1.67	1.27	1.05	0.96	0.91
0.90	11	1.88	1.43	1.18	1.08	1.02
1.00	9	2.09	1.59	1.31	1.20	1.13

2.8.1.4. Worked Example

Revisiting the worked example given earlier where the sample size was estimated to be 770 patients based a standardised width of 0.10. Now suppose the variance used in the calculations was estimated with 10 degrees of freedom. From Table 2.32 a high plausible value for the variance would have a standardised width of 0.16 – precision 60% worse than the standardised width upon which the sample size calculations were based.

2.8.1.5. Calculations Taking of Account the Imprecision of the Variance Used in the Sample Size Calculations

To account for the imprecision of the variance used in the sample size calculations for parallel group precision based trials, (2.7.4), and the results from superiority trials given earlier in this chapter, can be generalised to give the following formula

$$n_A \geq \frac{(r+1)s^2\left[tinv\left(0.5, \mathrm{m}, t_{1-\alpha/2, n_A(r+1)-2}\right)\right]^2}{rd^2}, \qquad (2.7.7)$$

where n_A is the least integer value for (2.7.7) to hold. This equation can in turn be rewritten as

$$0.5 \geq 1 - probt\left(\sqrt{\frac{rn_A d^2}{(r+1)s^2}}, m, t_{1-\alpha/2, n_A(r+1)-2}\right). \qquad (2.7.8)$$

Replacing the t-statistic with a Z-statistic gives one the following result

$$n_A = \frac{(r+1)s^2\left[tinv\left(0.5, \mathrm{m}, Z_{1-\alpha/2}\right)\right]^2}{rd^2} \qquad (2.7.9)$$

which allows a direct estimate for the sample size and gives an initial value for interations for (2.7.7).

Table 2.33 gives the sample sizes required for 95% confidence interval precision estimates and for a range of degrees of freedom, m, and standardised widths, w/s.

Table 2-33. Sample sizes for one group, n_A ($n_B=rn_A$) in a parallel group precision study for different standardised widths and degrees of freedom using (2.7.7) for a 5% level of significance. Sample sizes with "infinite" degrees of freedom are estimated from (2.7.3)

Degrees of	Standardised Widths					
Freedom	0.05	0.10	0.25	0.50	0.75	1.00
5	3434	860	139	36	17	10
10	3242	812	131	34	16	10
25	3138	786	127	33	16	10
50	3106	778	126	33	16	10
100	3090	774	125	33	15	10
∞	3075	770	125	32	15	9

Table 2.34 gives the multiplication factors for different levels of statistical significance. What is interesting to note is that for precision based studies the impact on the sample size is not as great as that for formally powered studies. This could be because power – and consequently the distribution under the alternative hypothesis – does not have to be considered for precision based studies. From Table 2.34 at most the impact is to increase the sample size by 12%.

Table 2-34. Multiplication factors for different levels of statistical significance

	Significance Level (α)			
m	0.010	0.025	0.050	0.100
5	1.122	1.120	1.117	1.114
10	1.056	1.055	1.054	1.053
25	1.021	1.021	1.021	1.021
50	1.010	1.010	1.010	1.010
75	1.007	1.007	1.007	1.007
100	1.005	1.005	1.005	1.005

2.8.1.6. Worked Example

Revisiting the worked example given earlier. For a standardised width of 0.10 and with 10 degrees of freedom about the variance estimate used in calculations the sample size 812 patients per arm compared to 770 calculated earlier. A relatively small increase in the sample size.

2.8.1.7. The Problem Reconsidered

The problem of uncertain variance estimates in the variability in context with precision based trials was discussed in detail by Grieve [1989, 1990, 1991] drawing on and commenting on the work of others [Beale, 1989; Day, 1988; Greenland, 1988; Kupper an Hafner, 1989]. Here, Grieve highlighted that for (2.7.2) or (2.7.3) one only had a 50% chance of achieving the desired precision as the variance half the time will be greater than that anticipated.

Grieve advocated the following result to work out the sample size

$$\Pr obability = probchi\left(\frac{w^2 r n_A(n_A(r+1)-2)}{(r+1)t^2_{1-\alpha\ 2,n_A(r+1)-2}\sigma^2}, n_A(r+1)-2\right), \qquad (2.7.10)$$

where $probchi(\bullet, n_A(r+1)-2)$ is a cumulative density distribution (using the same notation as SAS) of a χ^2 distribution on $n_A(r+1)-2$ degrees of freedom. The probability here is the probability of the confidence interval having the required precision for the variance estimated in the planned trial for a given sample size– it is not power. To estimate a sample size one has to iterate until a required level of probability is reached. This result compares with

$$n_A \geq \frac{(r+1)t^2_{1-\alpha\ 2,n_A(r+1)-2}\sigma^2}{rw^2}, \qquad (2.7.11)$$

which was given earlier or

$$n_A \geq \frac{(r+1)F_{1-\alpha\ 2,1,n_A(r+1)-2}\sigma^2}{rw^2}, \qquad (2.7.12)$$

which is (2.7.11) written in terms of a F-statistics.

Table 2-35. Sample sizes for one group, n_A (n_B=n_A) in a parallel group study for different standardised widths and probabilities for 95% confidence intervals for the precision estimates

	Probabilities			
δ	0.50	0.80	0.90	0.95
0.05	3076	3122	3146	3166
0.10	771	794	806	816
0.25	125	135	139	143
0.50	33	38	40	42
0.75	16	19	21	22
1.00	10	12	13	14

Table 2.35 gives sample sizes from (2.7.10) for given probabilities and required precisions. From visual inspection of this table and Table 2.31 (estimated from (2.7.11)) it is clear that for a probability of 0.5 (2.7.10) gives the same results as (2.7.4)

– allowing for a little rounding error. Also from inspection, it seems that to ensure a greater probability of having the required precision it does not require a great increase in the required sample size.

Table 2-36. Sample sizes for one group, n_A (n_B=n_A) in a parallel group precision study for different standardised widths, probabilities (p) and degrees of freedom using (2.7.13) for a 5% level of significance. Sample sizes with "infinite" degrees of freedom are estimated from (2.7.11)

p	Degrees of Freedom	Standardised Widths 0.05	0.10	0.25	0.50	0.75	1.00
0.50	5	3533	884	143	37	17	10
	10	3291	824	133	34	16	10
	25	3158	791	128	33	15	9
	50	3116	780	126	33	15	9
	100	3095	775	125	32	15	9
	∞	3075	770	124	32	15	9
0.80	5	6561	1642	264	67	31	18
	10	4976	1246	201	52	24	15
	25	4059	1071	165	43	21	13
	50	3710	930	152	40	19	12
	100	3499	878	144	39	19	12
	∞	3121	837	134	37	18	11
0.90	5	9544	2388	384	97	44	26
	10	6319	1582	255	66	31	18
	25	4667	1169	190	50	24	15
	50	4081	1024	167	45	22	14
	100	3737	938	155	42	21	13
	∞	3472	805	138	39	20	13
0.95	5	13417	3356	539	136	62	36
	10	7802	1953	315	81	37	22
	25	5262	1318	214	56	27	17
	50	4425	1110	182	49	24	15
	100	3950	993	164	45	22	14
	∞	3165	815	142	41	21	13

2.8.1.8. Allowing for the Imprecision in the Variance used in the Sample Size Calculations

In an elegant solution Grieve [1991] demonstrated that if there was prior uncertainty around the variance used in the precision sample size calculations then the probability of seeing the required precision for a given sample size can be calculated from

$$probability = probf\left(\frac{w^2 r n_A (n_A(r+1)-2)}{(r+1)t^2_{1-\alpha/2,n_A(r+1)-2}s^2}, n_A(r+1)-2, m\right), \qquad (2.7.13)$$

where $probf(\bullet, n_A(r+1)-2, m)$ is a cumulative density distribution (using the same notation as SAS) of a F distribution on $n_A(r+1)-2$ and m degrees of freedom. This

result was originally given without proof by Mood and Snedecor [1946]. Here s^2 is the estimate of the variance from a previous trial being used to plan the current trial, n_A is the sample size in the trial being planned, $n_A(r+1)-2$ the degrees of freedom of the variance in this trial and m is the degrees of freedom of s^2. To solve for n_A one iterates (2.7.13) until the appropriate probability is reached. The sample sizes from (2.7.13) are given in Table 2.36. For probabilities of 0.50 the table should be comparable to Table 2.33, however, the results from this table are a little conservative in comparison.

2.8.2. Cross-Over Trials

2.8.2.1. Sample Size Estimated Assuming the Population Variance to be Known

Similarly to the parallel group case one can solve (2.7.1) to give [Julious, 2004a]

$$n = \frac{2Z^2_{1-\alpha/2}\sigma^2_w}{w^2}, \qquad (2.7.14)$$

where n is the total sample size. If the population variance is to considered unknown in the statistical analysis (2.7.14) can be rewritten as [Julious, 2004a]

$$n \geq \frac{2t^2_{1-\alpha/2,n-2}\sigma^2_w}{w^2}, \qquad (2.7.15)$$

which can be solved iteratively. Alternatively as with parallel group trials the following formula could be used

$$0.5 \geq \Phi\left(\sqrt{\frac{nw^2}{2\sigma^2_w}} - t_{1-\alpha/2,n-2}\right). \qquad (2.7.16)$$

To allow for the Normal approximation, (2.7.14) can amended to have a correction factor [Guenther, 1981; Julious, 2004a]

$$n = \frac{2\sigma^2_w Z^2_{1-\alpha/2}}{w^2} + \frac{Z^2_{1-\alpha/2}}{2}. \qquad (2.7.17)$$

The following formula can be used assuming one wishes to have a 95% confidence interval precision estimates

$$n = \frac{8\sigma^2}{w^2}. \qquad (2.7.18)$$

Table 2.37 gives sample sizes using (2.7.16) for various standardised widths ($\delta = d/\sigma$). As with parallel group trials the quick formula slightly under estimates the sample size.

Table 2-37. Total sample sizes for a cross-over study for different standardised widths with 95% confidence intervals for the precision estimates

δ	n
0.10	771
0.20	195
0.30	88
0.40	51
0.50	34
0.60	24
0.70	19
0.80	15
0.90	13
1.00	11

2.8.2.2. Sensitivity Analysis About the Variance Used in the Sample Size Calculations

Table 2.38 gives sample sizes and precision estimates for high plausible values of the variance for different standardised widths. The inference drawn from this table is the same as for parallel group studies.

Table 2-38. Total sample sizes for a cross-over study for different standardised widths along with the precision for high plausible values for the variance

	Sample	Degrees of Freedom				
Difference	Size	5	10	25	50	100
0.10	771	0.21	0.16	0.13	0.12	0.11
0.20	195	0.42	0.32	0.26	0.24	0.23
0.30	88	0.63	0.48	0.39	0.36	0.34
0.40	51	0.84	0.64	0.52	0.48	0.45
0.50	34	1.04	0.80	0.65	0.60	0.57
0.60	24	1.25	0.96	0.78	0.72	0.68
0.70	19	1.46	1.12	0.92	0.84	0.79
0.80	15	1.67	1.27	1.05	0.96	0.91
0.90	13	1.88	1.43	1.18	1.08	1.02
1.00	11	2.09	1.59	1.31	1.20	1.13

2.8.2.3. Calculations Taking of Account the Imprecision of the Variance Used in the Sample Size Calculations

To account for the imprecision the results from parallel group trials can be generalised to give the following formula

$$n \geq \frac{2s_w^2\left[tinv\left(0.5, \mathrm{m}, t_{1-\alpha/2,n-2}\right)\right]^2}{d^2}. \qquad (2.7.17)$$

This equation can in turn be rewritten as

$$0.5 \geq probt\left(\sqrt{\frac{nd^2}{2s_w^2}}, m, t_{1-\alpha/2,n-2}\right). \qquad (2.7.18)$$

Replacing the t-statistic with a z-statistic gives one the following result

$$n = \frac{2s_w^2[tinv(0.5, \mathrm{m}, Z_{1-\alpha/2})]^2}{d^2}, \qquad (2.7.19)$$

which allows a direct estimate for the sample size and gives an initial value for interations for (2.7.17).

Table 2.39 gives the sample sizes required for 95% confidence interval precision estimates and for a range of degrees of freedom, m, and standardised widths, w/s. The multiplication factors for a cross-over trial are the same as those for a parallel group study given Table 2.34.

Table 2-39. Total sample sizes for cross-over precision study for different standardised widths and degrees of freedom using (2.7.18) for a 5% level of significance. Sample sizes with "infinite" degrees of freedom are estimated from (2.7.15)

Degrees of	Standardised Widths					
Freedom	0.05	0.10	0.25	0.50	0.75	1.00
5	3435	861	140	37	18	12
10	3243	813	133	35	18	11
25	3140	787	128	34	17	11
50	3107	779	127	34	17	11
100	3092	775	126	34	17	11
∞	3076	771	126	34	17	11

2.8.2.4. The Problem Reconsidered

The work of Grieve [1989, 1990, 1991] can be extended to cross-over trials such that the total sample size can be estimated from

$$\mathrm{Pr}obability = probchi\left(\frac{w^2 n(n-2)}{2rt^2_{1-\alpha/2,n-2}\sigma_w^2}, n-2\right). \qquad (2.7.20)$$

Table 2-40. Total sample sizes for a cross-over study for different standardised widths and probabilities for 95% confidence intervals for the precision estimates

	Probabilities			
δ	0.50	0.80	0.90	0.95
0.05	3075	3141	3175	3204
0.10	771	803	820	834
0.25	125	138	145	150
0.50	33	40	43	45
0.75	16	20	22	24
1.00	10	13	15	16

As with parallel group trials to estimate a sample size one has to iterate until a required level of probability is reached. Table 2.40 gives sample sizes from (2.7.20) for given probabilities and required precisions.

Table 2-41. Total sample sizes for a cross-over study for different standardised widths, probabilities (p) and degrees of freedom using (2.7.21) for a 5% level of significance. Sample sizes with "infinite" degrees of freedom are estimated from (2.7.20)

	Degrees of	Standardised Widths					
p	Freedom	0.05	0.10	0.25	0.50	0.75	1.00
0.5	5	3533	885	144	38	18	11
	10	3292	825	134	35	17	11
	25	3159	791	129	34	17	11
	50	3117	781	127	34	16	10
	100	3096	776	126	33	16	10
	∞	3075	771	125	33	16	10
0.8	5	6563	1643	266	69	32	20
	10	4977	1247	203	54	26	16
	25	4061	1019	167	45	23	14
	50	3713	933	154	42	21	14
	100	3502	881	147	41	21	14
	∞	3141	803	138	40	20	13
0.9	5	9546	2389	385	99	46	28
	10	6321	1584	257	68	32	20
	25	4670	1172	193	52	26	17
	50	4085	1027	171	48	24	16
	100	3742	943	159	45	23	15
	∞	3175	820	145	43	22	15
0.95	5	13418	3357	540	138	63	37
	10	7804	1955	317	83	39	24
	25	5265	1322	217	59	30	19
	50	4430	1114	186	52	27	18
	100	3956	999	169	49	25	17
	∞	3204	834	150	45	24	16

2.8.2.5. Allowing for the Imprecision in the Variance used in the Sample Size Calculations

Extending the parallel group result to allow for prior uncertainty around the variance used in the precision sample size calculations then the probability of seeing the required precision for a given sample size can be calculated from

$$\text{Probability} = probf\left(\frac{w^2 rn(n-2)}{(r+1)t^2_{1-\alpha,2,n-2}s^2_w}, n-2, m\right). \qquad (2.7.21)$$

To solve for n one iterates (2.7.21) to the appropriate probability is reached. Table 2.41 gives sample size tables using (2.7.21)

2.9. Design Considerations

2.9.1. Inclusion of Baselines or Covariates

In the analysis of the results of a clinical trial, the effects of treatment on the response of interest are often adjusted for predictive factors, such as demographic (like gender and age) or clinical covariates (such as baseline response), by fitting them concurrently with treatment. This section will concentrate on the case where baseline is the predictive covariate of interest (although the results are generalisable to other factors), the design is a parallel group design and an analysis of covariance, allowing for the baseline, is to be the final analysis. The CPMP have just issued notes for guidance on the design and analysis of studies with covariates [CPMP, 2003]

Frison and Pocock [1992] give a variance formula for various numbers of baseline measures

$$\text{Variance} = \sigma^2\left(1-\frac{p\rho^2}{1+(p-1)\rho}\right). \qquad (2.8.1)$$

Here, ρ is the Pearson correlation coefficient between observations – assuming compound symmetry - and p is the number of baseline measures taken per individual. From this equation a series of correction factors can be calculated [Machin, Campbell, Fayers, 1997] which give the variance reduction and consequent sample size reduction for different correlations and numbers of baselines. The assumption here is that there is balance between treatments and the baseline (or covariate) of interest. Any imbalance will increase the variance from (2.8.1), and consequent sample size [Senn, 1997]. With randomisation the imbalance should be minimised however.

From (2.8.1) it is clear that for fixed numbers of baseline measures the higher the correlation the greater the reduction in variance and consequent sample size. For example if three baseline measures were to be taken and the expected correlation between baseline and outcome is 0.5, the effect would be to reduce the variance to $0.6250\sigma^2$. However, for the same number of baseline measures if the expected correlation between baseline and outcome is 0.7 then the effect would be to reduce the variance to $0.3875\sigma^2$.

Table 2-42. Effect of number of baselines on the variance

Number of baselines	Variance
1	0.7500
2	0.6667
3	0.6250
4	0.6000
5	0.5833
6	0.5714

Another result from (2.8.1) is that for fixed correlation it seems that although there is incremental benefit with increasing numbers of baselines this incremental benefit asymptotes at 3 baselines. The results in Table 2.42 demonstrate this giving the correction factors for a fixed correlation between baseline and outcome of 0.50 and difference numbers of baseline measures.

The results of Frison and Pocock are a little simplistic – for example they assume that the within subject errors are independent [Senn, Stevens and Chaturvedi, 2000]. However, they do highlight the advantages of taking baselines in clinical trials.

The results in this sub-section demonstrate the importance, when estimating the sample size, to take the variance estimate from the full model where all covariates are present. It also highlights how, if one ignores baseline and covariate information when doing sample size calculations, one could potentially be overestimating the sample size. The variance allowing for covariates should be used in the sample size equations given in previous sections.

2.9.2. Post Dose Measures Summarised by Summary Statistics

Often in parallel group clinical trials, patients are followed up at multiple time points. Making use of all of the information obtained on a patient has the desirable property of increasing the precision for estimating the effects of treatment. Naturally as the precision is increased the variability is decreased and one consequently needs to study fewer patients in order to achieve a given power. Suppose one is interested in looking at the difference in the average of all of the post-dose measures

$$H_0 : \bar{\mu}_A = \bar{\mu}_B \text{ verses } H_1 : \bar{\mu}_A \neq \bar{\mu}_B ,$$

where $\bar{\mu}_A$ and $\bar{\mu}_B$ represent the means of the average the post-dose measures in the two treatment populations. Frison and Pocock [1992] explored several summary measures for this hypothesis. Often in a clinical trial where data are measured longitudinally, it is the rate of change of a particular endpoint which is of interest. For example in respiratory trials of chronic lung disease the hypothesis may focus on whether or not a treatment changes the annual decline in lung function. Diggle, Liang and Zeger [1994] describe the hypothesis for such a trial. However, the simplest approach of taking the summary measure as the simple average of the post-dose assessments for each subject and taking the average of these averages across treatments to obtain $\bar{\mu}_A$ and $\bar{\mu}_B$ is assumed to be the summary statistic used.

Assuming we have r post-dose measures and that the correlation between those measures is ρ the variance can be calculated as

$$\text{Variance} = \frac{\sigma^2[1+(r-1)\rho]}{r}, \qquad (2.8.2)$$

where σ^2 represents the variance of a given individual post-dose measurement.

When looking at (2.8.2) it seems that as the correlation between post-dose measures increases the variance increases so does the total sample size required. This is because, although it may seem counterintuitive, the advantage of taking additional measurements decreases as the correlation increases. This fact is due to how the total variance, σ^2, is constructed [Julious, 2000]

$$\sigma^2 = \sigma_b^2 + \sigma_w^2, \qquad (2.8.3)$$

where σ_w^2 is the within subject component of variation (as in cross-over trials) and σ_b^2 is the between subject component of variation.

It is important here to distinguish between the within- (intra-) subject and the between- (inter-) subject components of variation. The within-subject component of variation quantifies the expected variation among repeated measurements on the same individual. It is a compound of true variation in the individual. Whilst the between-subject component of variation, quantifies the expected variation of single measurements, from different individuals. If only one measurement is made per individual it is impossible to estimate σ_w^2 and σ_b^2 and consequently only the total variation, given in (2.8.3), can be estimated

If one knows the between-subject variance and the correlation between the measures the within-subject variance can be derived from

$$\sigma_w^2 = \left(\frac{1-\rho}{\rho}\right)\sigma_b^2. \qquad (2.8.4)$$

Therefore for known variance components of σ^2 and correlation between measures the variance that takes account of the number of post dose measures is defined as

$$\text{Variance} = \sigma_b^2 + \frac{\sigma_w^2}{r}. \qquad (2.8.5)$$

Thus, the formula (2.8.2) is actually quite intuitive. As for constant r the higher the correlation, from (2.8.4), the lower the within-subject variance and, from (2.8.5), the lower the total variance and consequent sample size. However, as ρ increases, and σ_w^2 falls, the effect of taking repeated measures diminishes as σ_w^2 already constitutes a small part of the overall variance.

Table 2-43. Effect of number of post dose measures on the variance

Number of post dose measures	Variance
1	1.0000
2	0.7500
3	0.6667
4	0.6250
5	0.6000
6	0.5833

Equation (2.8.2) also gives the incremental benefit of taking additional post dose measures for fixed correlation. Like with the number of baselines it seems that although there is incremental benefit with increasing numbers of post dose measures, the incremental benefit asymptotes at 4 post dose. The results in Table 2.43 demonstrate this giving the correction factors for a fixed correlation between post dose measures of 0.50 and difference numbers of post dose measures.

2.9.3. Inclusion of Baseline or Covariates as well as Post Dose Measures Summarised by Summary Statistics

As noted in the previous section further savings in sample size can be achieved by accounting for baseline as a covariate. Frison and Pocock [1992] define an additional variance measure to account for the baseline (or multiple baselines) as a covariate and difference numbers of post dose measures. Assuming there are p baseline visits and r post dose visits the variance is defined as

$$\text{Variance} = \sigma^2\left[\frac{1+(r-1)\rho}{r} - \frac{p\rho^2}{1+(p-1)\rho}\right]. \qquad (2.8.6)$$

2.10. Summary of Chapter 2

This chapter described in detail the standard calculations for sample size estimation for the most common types of clinical trial where the data are anticipated to take a Normal form. It was highlighted that one of the main assumptions in these calculations was with respect to the variance used in the calculations. This variance is usually a sample estimate that is estimated imprecisely and yet in calculations it is usually assumed fixed and known.

The chapter described a methodology to investigate the sensitivity of a study to the assumptions about the variance used in the sample size calculations. This sensitivity analysis was undertaken using the degrees of freedom for the sample variance and the chi-squared distribution to obtain a high plausible value for the variance. Determining the loss of power if the true variance was actually nearer to the high plausible value

assesses the sensitivity of the study. It was highlighted how if the variance was estimated with few degrees of freedom then the study was sensitive to assumptions about the variance. Recommendations were made as to how to optimise the variance estimate.

The chapter then went on to develop a methodology for estimating a sample size that accounts for the imprecision in the variance estimate – assessed through its degrees of freedom. It was demonstrated how having few degrees of freedom impacts on the sample size estimate. For example if one only had 10 degrees of freedom for the variance estimate and was designing a superiority study with 90% power and a two sided significance level of 5% one would require 30% more subjects to account for the imprecision in the variance. It was further shown that as the degrees of freedom increased, the precision of the variance also increased – impacting on the sample. Such that if one had 100 degrees of freedom or more one could use either the new results proposed in this chapter or standard calculations.

3. CHAPTER 3 - INFERENCE AND ANALYSIS OF CLINICAL TRIALS WITH BINARY DATA

3.1. Introduction

Binary outcomes are common endpoints in clinical trials and appear when the outcome of interest is a two point response variable such as presence/absence, alive/dead or yes/no. Researchers also often use cut-offs on continuous scales to dichotomise and form a binary outcome. In areas such as Quality of Life the cut-offs for some scales are well known and may define patients into various prognostic categories.

Note though that although these cut-offs may assist in interpretation, as will be discussed in Chapter 5, a lot of information is discarded if all the other categories are ignored, which can have a knock on in an increased sample size [Julious, George, Machin et al, 1997; Campbell, Julious and Altman, 1995].

For such a relatively straightforward response there are many issues associated with trials with binary primary outcome. Not least of these is how to summarise such data. Initially this chapter will describe four ways of summarising binary data:

1. Absolute risk reduction
2. Odds-ratio
3. Relative risk reduction
4. Number needed to treat

The chapter well describe the relative merits of each. These relative merits will be discussed in context with the difference types of trial that may be being conducted i.e. trials to assess superiority; equivalence and non-inferiority. The emphasis will be on assessing summary measures that can be generally used for all types of clinical trial.

For each summary statistic the methodology for the calculation of confidence intervals will be discussed. Obviously, when designing a trial *a priori* one should know both how one will summarise and analyse it. It will be highlighted how the limiting factor in making inference is the discrete nature of binary data - there are only a finite number of responses that may occur for a given sample size.

Particular emphasis will be given to the absolute risk reduction. This is because it is on this scale that the discrete (and bounded) nature of binary data is most apparent. Also the variance estimates for other summary measures are often dependent on the variance estimates on the proportional scale (estimated through Taylor's Series expansions).

The inference about the estimates discussed in this chapter will impact on the sample size discussions in Chapter 4. This is due to the sample size methodologies depending on the asymptotic properties of the parameters that the study is being powered on.

Note the objective of this chapter is not to give a definitive review of confidence interval methodology for binary responses but is intended as overview in the context of how the inference will affect sample size calculations.

3.2. Aims of the Chapter

The main issues to be covered in this chapter are as follows

- To describe the summary measures for assessing efficacy for a binary response and how these measures are impacted by different clinical trial objectives.
- To review the methodologies for the calculation of a confidence interval for a single binary response.
- To describe the methodologies for the calculations of confidence intervals for the different summary measures for a parallel group trial.
- To explore the asymptotic (and other) properties for the estimates of response and the variance about these responses.
- To make recommendations as to the most appropriate summary measures for binary responses.

3.3. Absolute Risk Reduction

For a clinical trial where the primary outcome is a binary response the data may take the form summarised in Table 3.1 where: p_A and p_B are the responses anticipated on treatment A and B respectively; $\bar{p}$ the average response across treatments; n_A and n_B are the sample sizes in each treatment group and N is the total sample size.

Table 3-1. Summary table for a clinical trial with a binary outcome

	Outcome		
Treatment	0	1	Sample Size
A	$1-p_A$	p_A	n_A
B	$1-p_B$	p_B	n_B
Overall Response	$1-\bar{p}$	$\bar{p}=(p_A+p_B)/2$	$N=n_A+n_B$

The absolute risk reduction is probably the simplest way of summarising binary data. One simply takes the risk of the event for each treatment, p_A and p_B, and takes the absolute difference of these $p_A - p_B$.

One drawback of working on the proportional scale is that the difference is bounded by (-1, 1). This bounding can adversely affect inference – especially when a response is near one of the bounds. The affect of bounding of the proportional scale will be discussed throughout the chapter.

3.3.1. Absolute Risk Reduction and Clinical Trials

This sub-section will walk through how a study, in terms of the null and alternative hypothesis, would be designed when the absolute risk reduction (or improvement) is used as the assessment of efficacy.

3.3.1.1. Superiority Trials

When designing a superiority trial in terms of absolute effects, π_A, π_B and $\pi_A - \pi_B$, the null hypothesis (H_0) and alternative hypothesis (H_1) are defined as

$H_0 : \pi_A = \pi_B$ or $H_0 : \pi_A - \pi_B = 0$,

and

$H_1 : \pi_A - \pi_B = d$,

where here d is some pre-define treatment effect that used in sample size calculations.

3.3.1.2. Equivalence Trials

The null and alternative hypotheses for an equivalence trial, on a proportional scale, take the form

$H_0 : \pi_A - \pi_B \geq d$ or $\pi_A - \pi_B \leq -d$,

$H_1 : -\pi < p_A - p_B < \pi$.

As discussed in Chapter 2, this is an example of an intersection-union test (IUT), in which the null hypothesis is expressed as a union and the alternative as an intersection. This is operationally the same as constructing a (1-2α)100% confidence interval and to conclude equivalence provided that the ends of the confidence interval fall completely within the interval $(-d,+d)$.

3.3.1.3. Non-Inferiority Trials

On the proportional scale the null and alternative hypotheses take the form

$H_0 : \pi_A - \pi_B \leq -d$,

$H_1 : \pi_A - \pi_B > -d$.

In order to conclude non-inferiority, one needs to reject the null hypothesis. In practice, this is the same as constructing a (1-2α)100% confidence interval and concluding non-inferiority provided that the lower end of this confidence interval is above –d on the proportional scale.

3.3.1.4. Choice of Non-Inferiority Limit

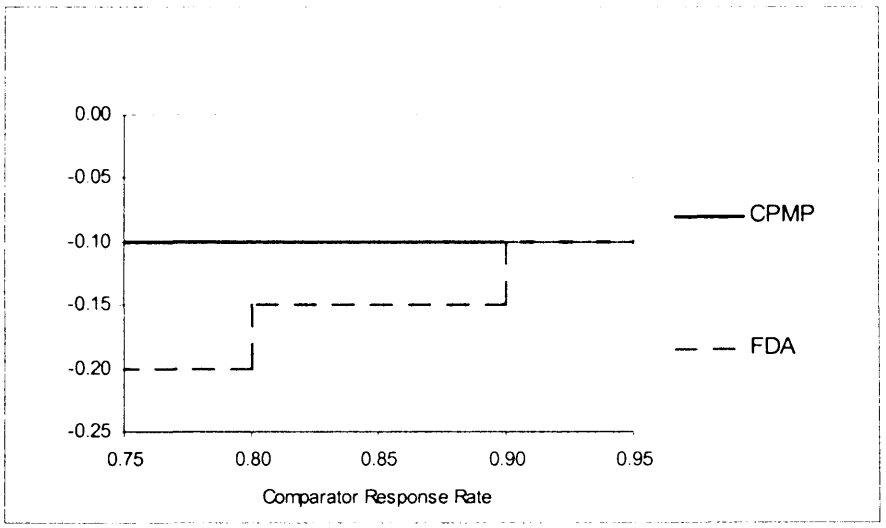
The choice of non-inferiority (and equivalence) limits was discussed generally in Chapter 1. However, it is worth re-interrogating this issue for binary data, as it is one of the few areas where there is hard regulatory guidance on the issue. The guidance is for the antimicrobial therapeutic area where active controlled trials are the norm although the issues raised are generic to other therapeutic areas.

Table 3-2. Non-inferiority margins for different control response rates

	Non-inferiority Margin	
Response Rate	FDA	CPMP
≥90	-10%	10%
80-89%	-15%	10%
70-79%	-20%	10%

Table 3.2 gives the non-inferiority margins for different response rates as recommended by CPMP [2004] and FDA [1992].

Figure 3-1. Graphical illustration of CPMP and FDA non-inferiority limits



What is evident from Table 3.2 is that whilst the CPMP recommend a flat equivalence margin the FDA currently recommends a step function according to the anticipated control response rate. Table 3.2 is also figuratively described in Figure 3.1.

Having different regulatory guidance has two practical problems first there is the need to design a study that sufficiently meets the regulatory requirements of two regions. The second is how to design a study – should it be based on the more stringent guidance?

3.3.2. Calculation of Confidence Intervals

3.3.2.1. Single Proportion

This chapter will now discuss the methodology for the calculation of confidence intervals, and corresponding inference. The rationale for this discussion is that the inference for a single proportion generalises both to the absolute difference in two proportions and other binary summaries such as odds-ratios and relative risks. Inference about a single proportion is also important when designing a clinical trial, as often one only has data for just one arm (usually the control arm). From this one arm a response rate is inferred for the other (usually investigative arm). Thus, it is the imprecision of a single proportion to which a study is sensitive as will be discussed in Chapter 4.

Table 3-3. Anticipated frequency distributions for different population responses

n	π =0.4	π =0.5	π =0.6
0	0.00004	0.00000	0.00000
1	0.00049	0.00002	0.00000
2	0.00309	0.00018	0.00000
3	0.01235	0.00109	0.00004
4	0.03499	0.00462	0.00027
5	0.07465	0.01479	0.00129
6	0.12441	0.03696	0.00485
7	0.16588	0.07393	0.01456
8	0.17971	0.12013	0.03550
9	0.15974	0.16018	0.07099
10	0.11714	0.17620	0.11714
11	0.07099	0.16018	0.15974
12	0.03550	0.12013	0.17971
13	0.01456	0.07393	0.16588
14	0.00485	0.03696	0.12441
15	0.00129	0.01479	0.07465
16	0.00027	0.00462	0.03499
17	0.00004	0.00109	0.01235
18	0.00000	0.00018	0.00309
19	0.00000	0.00002	0.00049
20	0.00000	0.00000	0.00004

Inference for a single proportion generalises to other parameters such as relative risk and odds-ratios as the standard errors for these parameters may be estimated from the standard error of a single proportion through approximations using the delta method [Armitage and Berry, 1987].

To consider the inferences about a single proportion one must go back to basics. Table 3.3 gives an illustration of the anticipated frequency distribution for different proportions for a fixed sample size of 20. This is known as a binomial distribution. From this table two major points are evident. The binomial distribution although uni-modal is only symmetric when $\pi = 0.5$. For any other value the distribution is skewed. The other point is the discrete nature of the distribution. As the sample size tends to infinity the binomial distribution tends towards the Normal [Kendall and Stuart, 1977] however, the discreteness of the distribution may question the robustness of this assumption for small sample sizes. This assumption will be interrogated throughout the chapter.

3.3.2.2. Normal Approximation

Under the Normal approximation the confidence interval for a single proportion is defined as

$$p \pm Z_{1-\alpha/2} se(p), \tag{3.2.1}$$

where $se(p) = \sqrt{p(1-p)/n}$. (3.2.2)

where p is the estimated response from the trial. This method is referred to as the Wald method [Newcombe, 1998a]. An empirical investigation of the asymptotic assumptions is given in Figures 3.2 to 3.4. This was done through simulation in SAS [1990]. For each sample size 10,000 simulations were undertaken assuming the population prevalence rate was 60%. One thing to highlight though in the simulations is that except for the large sample sizes there was a degree of redundancy in these simulations due to the discrete (and very finite) distribution being sampled from.

Figure 3.2 gives the Normal probability plot for the simulated proportions. As one can see from this plot the simulated proportions do not deviate greatly from the approximation to Normal for all the sample sizes. For each simulation the following was also calculated

$$(n-1)\frac{p(1-p)/n}{\pi(1-\pi)/n} = (n-1)\frac{p(1-p)}{\pi(1-\pi)},$$

where p is the estimated response from the simulation and π is the population prevalence (p=0.60) from which each simulation was drawn.

Figure 3-2. Normal probability plots for different sample sizes for a response (π=0.6) sampled from a binomial distribution

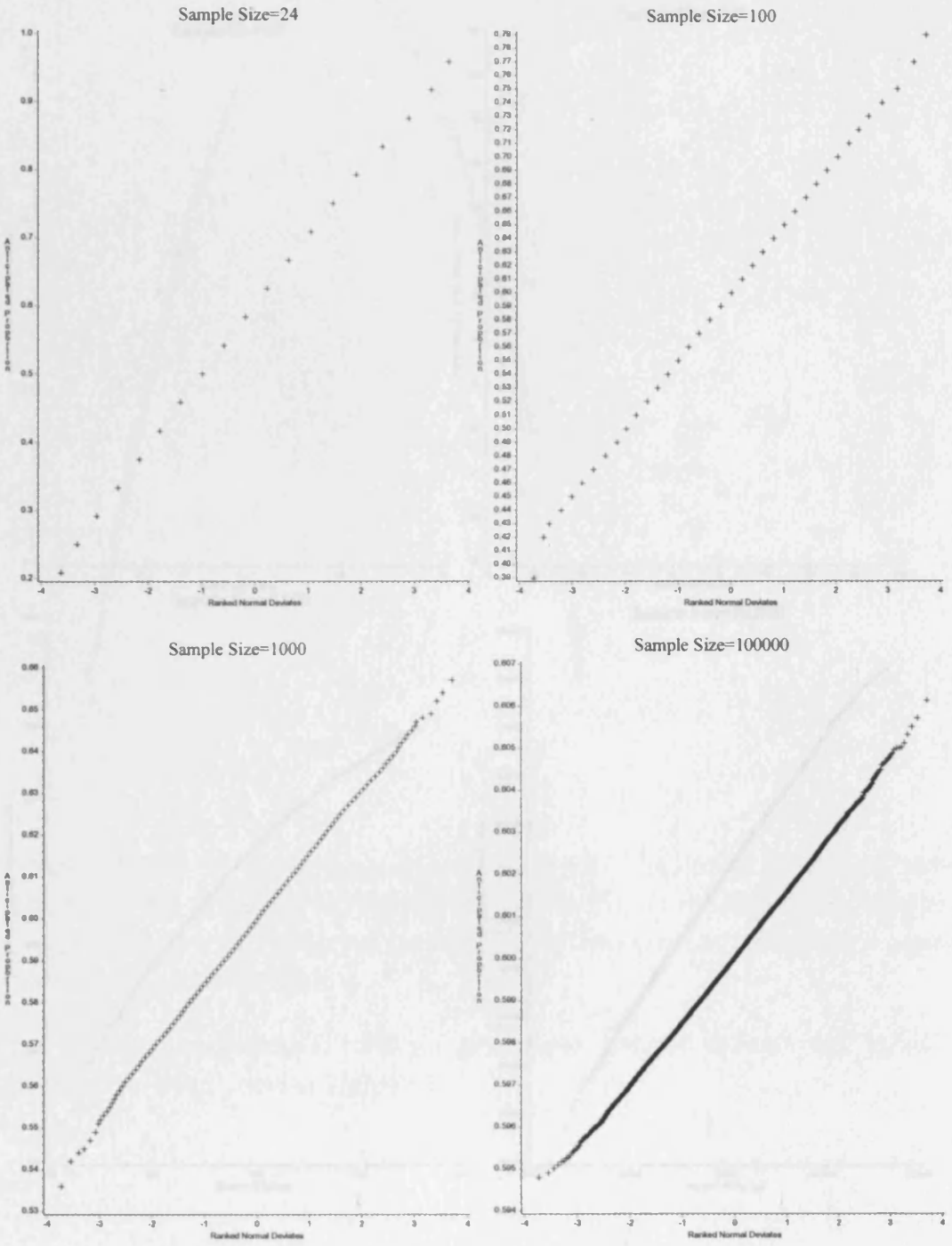
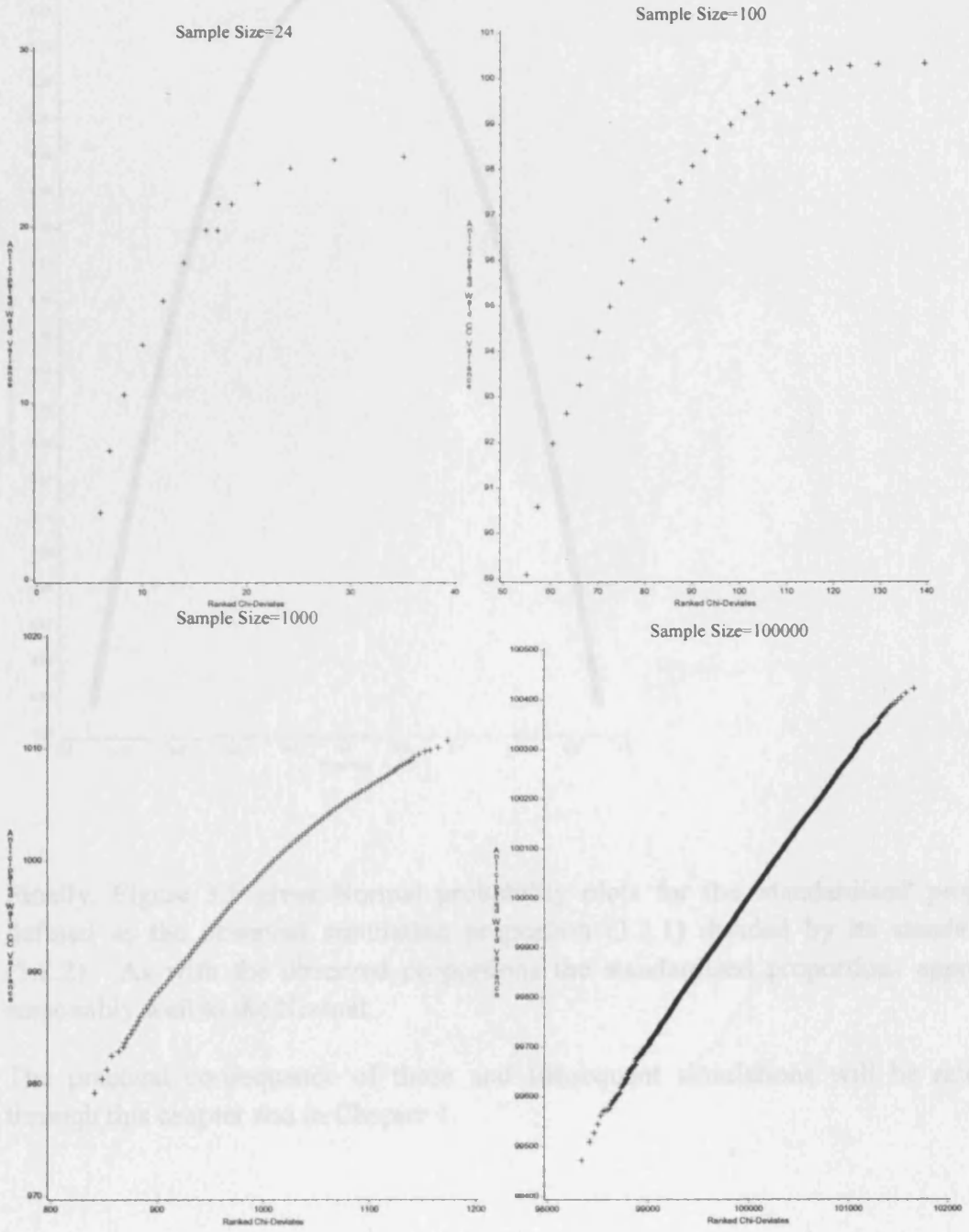


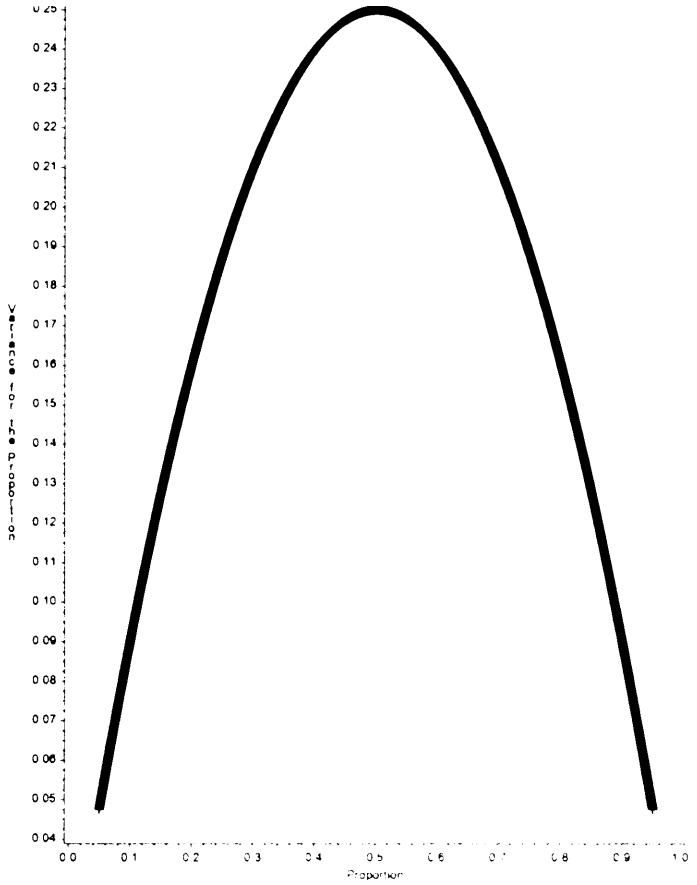
Figure 3.3 gives chi-probability plots for the ratio of the simulated sample over the population variance for different sample sizes. What is evident from this figure is that the approximation to the chi-distribution is quite weak for small sample sizes but improves the larger the trial.

Figure 3-3. Chi-probability plots for different sample sizes for a Wald variance (π=0.6) sampled from a binomial distribution



A rationale for Figure 3.3 may be gained from Figure 3.4. This gives a plot of the variance, p(1-p), against an observed proportion, p, for different values of p. One can see from this figure that with p in the centre of the range (0.3 to 0.7) the variance is quite stable. It is only as p approaches the extreme of the range (0 or 1) that the variance differs markedly. Thus, if p is towards the middle of the range and is imprecisely estimated (i.e. from a small sample size) the consequent plausible range for the variance would be comparatively narrow - less than expected if the sample variance was for continuous data and followed a chi-squared distribution.

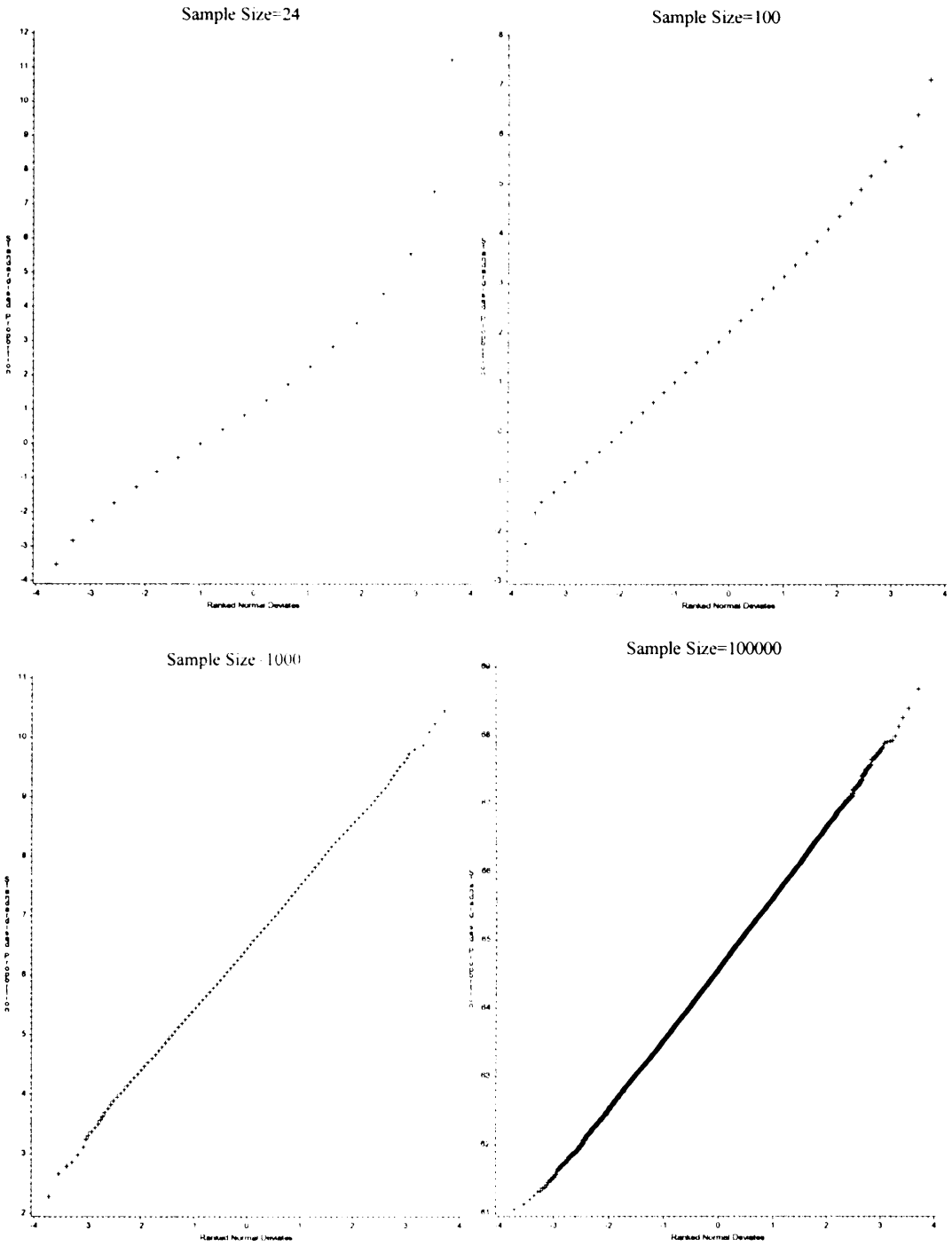
Figure 3-4. Plot of the variance of a proportional response for different responses



Finally, Figure 3.5 gives Normal probability plots for the 'standardised' proportions defined as the observed simulation proportion (3.2.1) divided by its standard error (3.2.2). As with the observed proportions the standardised proportions approximate reasonably well to the Normal.

The practical consequence of these and subsequent simulations will be referred to through this chapter and in Chapter 4.

Figure 3-5. Normal probability plots for different sample sizes for a standardised proportional response (π=0.6) sampled from a binomial distribution



3.3.2.3. Normal Approximation with Continuity Correction

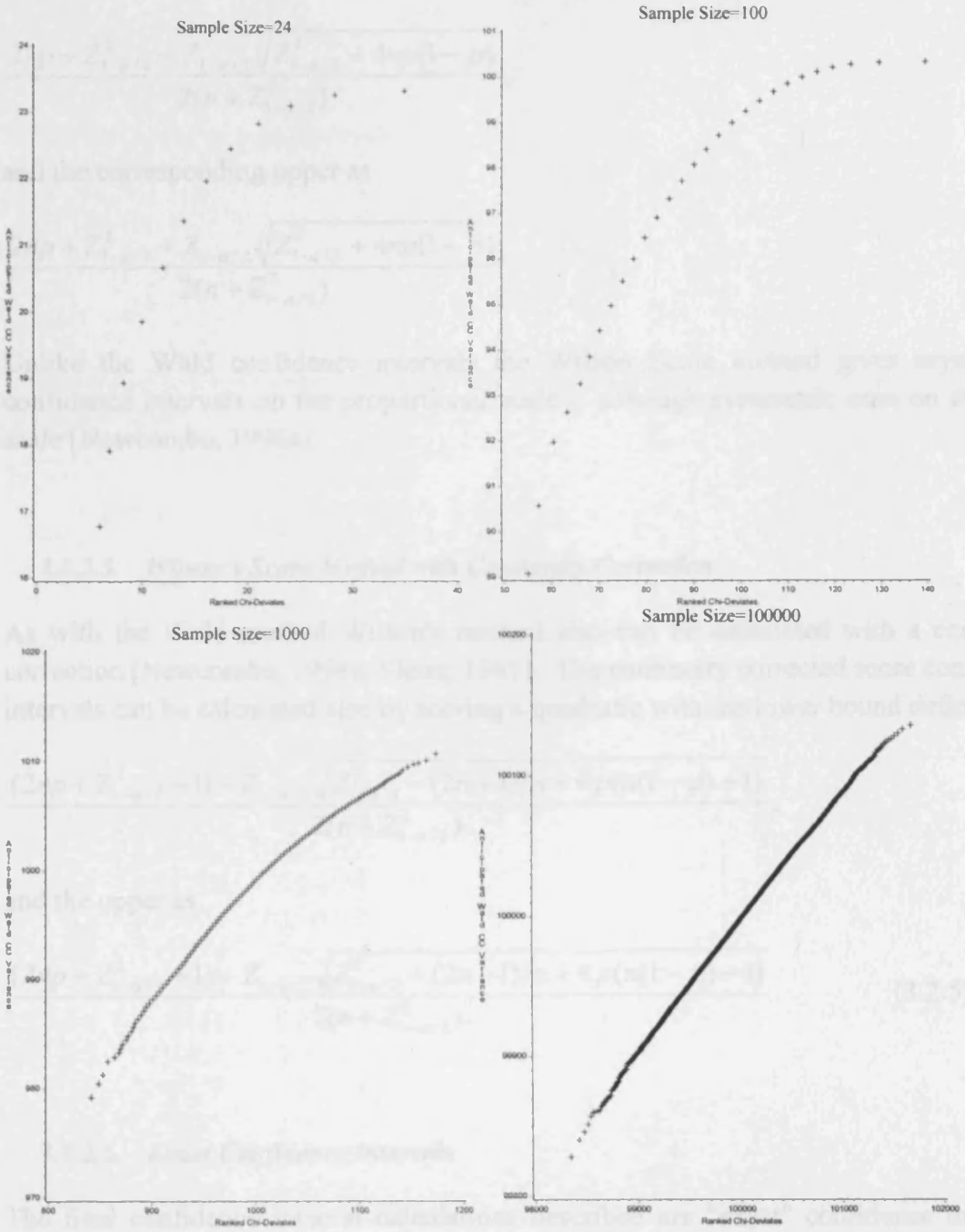
To approximate better to the Normal quantiles one can add a continuity correction to the confidence interval through addition to the right hand side of (3.2.2) of $1/(2n)$ [Newcombe, 1998a; Fleiss, 1981]. Thus, the standard error of p can be rewritten as

$$se(p) = \sqrt{\frac{p(1-p)}{n}} + \frac{1}{2n}, \qquad (3.2.3)$$

and used in (3.2.1) to construct a confidence interval.

Figure 3.6 gives a chi-probability plot for the variance defined from (3.2.3). As with the probability plots for none continuity corrected variances the evidence tends to suggest that the approximation to the chi-squared distribution is weak if the sample size is small.

Figure 3-6. Chi-probability plots for different sample sizes for continuity corrected Wald variance (π=0.6) sampled from a binomial Distribution



3.3.2.4. *Wilson's Score Method*

An alternative method for calculating confidence intervals is the score method [Newcombe, 1998] first introduced by Wilson [1927]. To calculate the confidence interval one solves the following quadratic

$$\frac{2np + Z^2_{1-\alpha/2} \pm Z_{1-\alpha/2}\sqrt{Z^2_{1-\alpha/2} + 4np(1-p)}}{2(n + Z^2_{1-\alpha/2})}, \qquad (3.2.4)$$

such that the lower point of the confidence interval is defined as

$$\frac{2np + Z^2_{1-\alpha/2} - Z_{1-\alpha/2}\sqrt{Z^2_{1-\alpha/2} + 4np(1-p)}}{2(n + Z^2_{1-\alpha/2})},$$

and the corresponding upper as

$$\frac{2np + Z^2_{1-\alpha/2} + Z_{1-\alpha/2}\sqrt{Z^2_{1-\alpha/2} + 4np(1-p)}}{2(n + Z^2_{1-\alpha/2})}.$$

Unlike the Wald confidence intervals the Wilson Score method gives asymmetric confidence intervals on the proportional scale – although symmetric ones on the logit scale [Newcombe, 1998a].

3.3.2.5. *Wilson's Score Method with Continuity Correction*

As with the Wald method Wilson's method also can be calculated with a continuity correction [Newcombe, 1998a; Fleiss, 1981]. The continuity corrected score confidence intervals can be calculated also by solving a quadratic with the lower bound defined as

$$\frac{(2np + Z^2_{1-\alpha/2} - 1) - Z_{1-\alpha/2}\sqrt{Z^2_{1-\alpha/2} - (2n+1)\ n + 4p(n(1-p)+1)}}{2(n + Z^2_{1-\alpha/2})},$$

and the upper as

$$\frac{(2np + Z^2_{1-\alpha/2} + 1) + Z_{1-\alpha/2}\sqrt{Z^2_{1-\alpha/2} + (2n-1)\ n + 4p(n(1-p)-1)}}{2(n + Z^2_{1-\alpha/2})}. \qquad (3.2.5)$$

3.3.2.6. *Exact Confidence Intervals*

The final confidence interval calculations described are "exact" confidence intervals, also known as Clopper-Pearson confidence intervals [Clopper and Pearson, 1934]. These confidence intervals are calculated by summing each of the tail probabilities from the binomial distribution, given the observed number of cases (k) for the sample size (n).

Defining the individual cell probabilities as

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)}, \qquad (3.2.6)$$

the lower limit of the confidence interval is defined as the lowest cumulative value of p such that the lower tail area of the distribution is no more than $\alpha/2$. Likewise the upper limit is calculated as the point where the cumulative distribution exceeds $1-\alpha/2$. Formally, the lower point of a confidence interval is defined as

$$\sum_{i=0}^{k} \binom{n}{i} p_L^i (1-p_L)^{(n-i)} < \alpha/2,$$

whilst the upper point is defined as,

$$\sum_{k=0}^{k} \binom{n}{i} p_U^i (1-p_U)^{(n-i)} > 1-\alpha/2. \qquad (3.2.7)$$

Table 3-4. Frequency and cumulative frequency distributions for a sample size of 20 and response of 0.60

n	Probability	Cumulative Probability
0	0.00000	0.00000
1	0.00000	0.00000
2	0.00000	0.00001
3	0.00004	0.00005
4	0.00027	0.00032
5	0.00129	0.00161
6	0.00485	0.00647
7	0.01456	0.02103
8	0.03550	0.05653
9	0.07099	0.12752
10	0.11714	0.24466
11	0.15974	0.40440
12	0.17971	0.58411
13	0.16588	0.74999
14	0.12441	0.87440
15	0.07465	0.94905
16	0.03499	0.98404
17	0.01235	0.99639
18	0.00309	0.99948
19	0.00049	0.99996

Table 3.4 illustrates the effect of the sample size on the confidence interval calculation. In the example, the sample size was 20 with 12 subjects a prevalence of 0.60. From the

column of cumulative probabilities the 95% confidence intervals is thus estimated as 0.35 to 0.80.

The link between the F distribution and the binomial distribution can also be used to calculate exact confidence intervals such that the lower bound is defined as [Agresti and Coull, 1998; Anderson and Burnstein 1967 and 1968; Casella, 1986; Crow, 1956; Daly, 1992; Edwardes, 1998; Ghosh, 1979; Blyth, 1983; Korn, 1986; Sterne, 1945; Vollset, 1993]

$$\frac{k}{k+(n-k+1)F_{1-\alpha\ 2,2n-2k+2,2k}},$$

and upper bounds are defined as

$$\frac{k+1}{k+1+(n-k)\ F_{1-\alpha\ 2,2k+2,2n-2k}}. \qquad (3.2.8)$$

It is interesting how using tails of the F-distribution has become the "norm" for the calculation of exact confidence intervals when a more straightforward calculation can be made through a using of tails of a Beta distribution [Julious, 2005b; Daly, 1992; Johnson and Kotz 1994a, 1994b, 1994c; Reiczigel, 2003; Newcombe, 1998]. To demonstrate this one needs to go back to basics and consider the distribution functions of the F-distribution

$$p_F(x;a,b)=\frac{a^{a\ 2}b^{b\ 2}\Gamma(a\ 2+b\ 2)}{\Gamma(a\ 2)\Gamma(a\ 2)}\int_0^x\frac{t^{(a-2)\ 2}}{(at+b)^{(a+b)\ 2}}dt,$$

and the Beta distribution

$$p_B(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x t^{a-1}(1-t)^{b-1}dt.$$

From inspection of these it is apparent that the following relationship holds [Daly, 1992; Julious 2005b]

$$F_{P,a,b}=\frac{bB_{P,a\ 2,b\ 2}}{a\left(1-B_{P,a\ 2,b\ 2}\right)}.$$

Which when substituted back in (3.2.8) gives lower bound defined as

$$1-BETAINV\left(1-\alpha/2,n-k+1,k\right),$$

and upper as

$$BETAINV\left(1-\alpha/2,k+1,n-k\right). \qquad (3.2.9).$$

A more straightforward nomenclature. The fact that the calculations using the Beta distribution are not more commonly used one imagines to be simply a function of the

fact that F-distribution are more readily available compared to the Beta distribution. This should not continue to the present day as $BETAINV(\bullet)$ is a function in packages such as SAS (even Excel has a similar function) and so the confidence interval calculation is operationally straightforward.

It should be noted here, however, that just because the "exact" confidence intervals have this flag exact does not make them truly "exact". This is because one only knows the true coverage probabilities of these intervals if one knows the true population rate [Clayton and Mills, 1993].

3.3.2.7. Comparison of the Different Methods

Table 3.5 give a comparison of the different methods of calculating the confidence intervals for different sizes. For large(ish) sample sizes (100+) there is pretty good agreement between the different methods. Even for a relatively small sample size of 50 to two decimal places the agreement is relatively good. It is with small sample sizes that the differences are more apparent with both the Wald confidence intervals giving upper bounds greater than one in these examples. As one would expect using the tails of the Beta distribution matches the summing of tails of the binomial distribution.

The selection of the most appropriate confidence interval depends on the objective for which the calculation will be used. If one wishes to describe a plausible range of values for a given confidence interval then a generic solution would be to quote Wilson (continuity or non continuity corrected) confidence intervals. If, however, the objective is to rule out a certain value with a pre-specified probability (say for a non-inferiority study) then the exact confidence intervals are probably the optimal solution as these will provide a lower bound (say) from which one can guarantee to be greater than the pre-specified value with a probability less than or equal to some P-value.

3.3.2.8. Difference in Two Proportions

The methods for the calculation of confidence intervals for the difference in independent proportions are similar to those for a single proportion and will be briefly discussed now.

Table 3-5. A comparison of the different methods for calculating confidence intervals

	Sample	Bionomial		Beta		Wald		Wald (with CC)		Wilson		Wilson (with CC)	
Proportion	Size	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
0.60	20	0.3500	0.8000	0.3605	0.8088	0.3853	0.8147	0.3603	0.8397	0.3866	0.7812	0.3641	0.8007
	50	0.4400	0.7400	0.4518	0.7359	0.4642	0.7358	0.4542	0.7458	0.4618	0.7239	0.4520	0.7327
	100	0.4900	0.6900	0.4972	0.6967	0.5040	0.6960	0.4990	0.7010	0.5020	0.6906	0.4970	0.6952
	500	0.5440	0.6420	0.5556	0.6432	0.5571	0.6429	0.5561	0.6439	0.5565	0.6420	0.5554	0.6430
	1000	0.5690	0.6300	0.5689	0.6305	0.5696	0.6304	0.5691	0.6309	0.5693	0.6299	0.5688	0.6304
	10000	0.5904	0.6096	0.5903	0.6096	0.5904	0.6096	0.5903	0.6097	0.5904	0.6096	0.5903	0.6096
0.75	20	0.5000	0.9000	0.5090	0.9134	0.5602	0.9398	0.5352	0.9648	0.5313	0.8881	0.5059	0.9046
	50	0.6200	0.8600	0.6074	0.8616	0.6300	0.8700	0.6200	0.8800	0.6151	0.8492	0.6045	0.8571
	100	0.6500	0.8300	0.6534	0.8312	0.6651	0.8349	0.6601	0.8399	0.6570	0.8245	0.6516	0.8288
	500	0.7100	0.7980	0.7096	0.7874	0.7120	0.7880	0.7110	0.7890	0.7102	0.7860	0.7092	0.7869
	1000	0.7220	0.7770	0.7219	0.7766	0.7232	0.7768	0.7227	0.7773	0.7222	0.7758	0.7217	0.7763
	10000	0.7414	0.7584	0.7414	0.7585	0.7415	0.7585	0.7415	0.7585	0.7414	0.7584	0.7414	0.7584
0.90	20	0.7000	1.0000	0.6830	0.9877	0.7685	1.0315	0.7435	1.0565	0.6990	0.9721	0.6687	0.9832
	50	0.8000	0.9800	0.7819	0.9667	0.8168	0.9832	0.8068	0.9932	0.7864	0.9565	0.7741	0.9627
	100	0.8300	0.9500	0.8238	0.9510	0.8412	0.9588	0.8362	0.9638	0.8256	0.9448	0.8196	0.9484
	500	0.8700	0.9260	0.8703	0.9249	0.8737	0.9263	0.8727	0.9273	0.8706	0.9233	0.8695	0.9242
	1000	0.8800	0.9180	0.8797	0.9179	0.8814	0.9186	0.8809	0.9191	0.8798	0.9171	0.8793	0.9175
	10000	0.8940	0.9058	0.8940	0.9058	0.8941	0.9059	0.8941	0.9059	0.8940	0.9057	0.8939	0.9058

3.3.2.9. *Normal Approximation*

Under Normal approximation the confidence interval for the difference in proportions is defined as

$$p_A - p_B \pm Z_{1-\alpha/2} se(p_A - p_B), \tag{3.2.10}$$

where $se(p_A - p_B) = \sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}$. (3.2.11)

This method is referred to as the Wald method [Newcombe 1998b].

3.3.2.10. *Normal Approximation with Continuity Correction*

One can add a continuity correction to the confidence interval through addition to the right hand side of (3.2.10) of $(1/n_A + 1/n_B)/2$ [Newcombe 1998b; Fleiss, 1981] such that it takes the form

$$p_A - p_B \pm Z_{1-\alpha/2} se(p_A - p_B) + (1/n_A + 1/n_B)/2. \tag{3.2.12}$$

3.3.2.11. *Wilson's Score Method*

The Wilson Score confidence intervals are derived from the following, for the lower [Newcombe, 1998b]

$$L = p_A - p_B - \delta,$$

and upper bounds

$$U = p_A - p_B - \varepsilon, \tag{3.2.13}$$

where

$$\delta = Z_{1-\alpha/2}\sqrt{l_A(1-l_A)/n_A + u_B(1-u_B)/n_B} \text{ and } \varepsilon = Z_{1-\alpha/2}\sqrt{u_A(1-u_A)/n_A + l_B(1-l_B)/n_B}.$$

Here l_A and u_A are the lower and upper bounds for p_A and l_B and u_B are the lower and upper bounds for p_B obtained from (3.2.4)

3.3.2.12. *Wilson's Score Method with Continuity Correction*

The Wilson Score continuity corrected confidence intervals are derived as per (3.2.13) but with l_A, u_A, l_B and u_B (the lower and upper bounds for p_A and p_B respectively) obtained from (3.2.5) instead of (3.2.4).

3.3.2.13. Exact Confidence Intervals

For two independent proportions p_A (r_A events in n_A subjects) and p_B (r_B events in n_B subjects) the probability function for their difference $\theta = p_A - p_B$ can be expressed in terms of θ and a nuisance parameter p_B [Agresti.2003; Agresti and Min, 2001; Newcombe 1998b]

$$f(r_A, r_B; n_A, n_B, \theta, p_B) = \binom{n_A}{r_A}(\theta + p_B)^{r_A}(1 - \theta - p_B)^{(n_A - r_A)}\binom{n_B}{r_B}p_B^{r_B}(1 - p_B)^{(n_B - r_B)}. \quad (3.2.14)$$

To obtain the lower and upper bounds for the 95% confidence (3.2.14) could be used through iteration to obtain the 2.5th and 97.5th percentile. In truth this confidence interval is not exact in the strictest sense, more a permutation type "exact" confidence interval [Santer and Snell, 1980].

3.3.2.14. Comparison of the Different Methods

To illustrate the different methods of confidence interval estimation, Table 3.6 was constructed to give a comparison of the different methods for fixed proportions of p_A=0.40 and p_B=0.60 ($p_A - p_B = -0.20$) for different sample sizes. What this table highlights is that if there is sample of 50 per group there is little separation between the different methods. Even for a sample size of 20 per group the agreement is relatively good.

As with the comparison of the methods earlier in this chapter for a single proportion the method to use for a small sample size depends on the rational for the calculation. For consistency with the single proportion case a recommendation would be to use the exact calculations if there is a value one wished to be ruled out with certain probability and Wilson (continuity or non continuity corrected) confidence intervals if a plausible range of values needs to be described with an appropriate coverage.

Table 3-6. A Comparison of the different Methods for Calculating Confidence Intervals for Two Proportions of P_A=0.40 and P_B=0.60

Sample	Exact		Wald		Wald (with CC)		Wilson		Wilson (with CC)	
Size	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
10	-0.20	0.70	-0.23	0.63	-0.33	0.73	-0.21	0.53	-0.19	0.50
20	-0.10	0.55	-0.10	0.50	-0.15	0.55	-0.10	0.46	-0.10	0.45
50	0.00	0.38	0.01	0.39	-0.01	0.41	0.00	0.38	0.00	0.37
100	0.06	0.33	0.06	0.34	0.05	0.35	0.06	0.33	0.06	0.33
500	0.14	0.26	0.14	0.26	0.14	0.26	0.14	0.26	0.14	0.26
1000	0.16	0.24	0.16	0.24	0.16	0.24	0.16	0.24	0.16	0.24

3.4. Number Needed to Treat

The number needed to treat (NNT) is a parameter that has been advocated for binary type data. This is because it supposedly gives an estimated effect in terms of a number of subjects as opposed to more "abstract" probabilities and risks which do not relate so directly to day to day practice [Cook and Sackett, 1995].

The number needed to treat (NNT) is defined as

$$NNT = \frac{1}{p_A - p_B}, \qquad (3.3.1)$$

where p_A and p_B are the proportion of subjects expected to have an event on regimens A and B respectively. It is thus the reciprocal of the proportional difference between treatments. It is interpreted as the number of subjects that on average would need to be treated on the investigative therapy to prevent one event that otherwise would have occurred had the control therapy been given. For example an NNT of 10 means that on average after treating each of 10 subjects one would expect to have prevented one event. Hence, it is argued that the two proportions, p_A and p_B, have been reduced into a single parameter to provide an estimate of the treatment effect.

The published literature indicates there has been an increase in the use of NNT, as evidenced in the British Medical Journal where articles quoting the parameter increased from one article in 1994 to a peak of 34 in 1998. This has led to some discussion as to its merits [Grieve, 2003; Hutton, 2000; Julious, 2002, 2005c; Smeeth, Haines and Ebrahim, 1999].

3.4.1. Number Needed to Treat and Clinical Trials

There are a number of issues with using NNT in clinical trials as will be now highlighted.

3.4.1.1. Superiority Trials

When designing a superiority trial, one should plan the study with reference to a null (H_0) and alternative hypothesis (H_1). As discussed earlier in this chapter when one is thinking in terms of p_A, p_B and $p_A - p_B$ the null hypothesis is easy to define

$H_0 : \pi_A = \pi_B$ or $H_0 : \pi_A - \pi_B = 0$,

and

$H_1 : \pi_A - \pi_B = d$,

where d is some pre-define treatment effect. However what is the null hypothesis when planning to summarise the trial using NNT [Hutton; Julious, 2002, 2005c]

$H_0 : NNT = \infty$?

3.4.1.2. Non-Inferiority Trials

NNT has also been advocated to summarise non-inferiority trials and equivalence trials [Bender, 2001]. On the face of it for non-inferiority trials there may be some merit to the proposal. Remember that on the proportional scale the null and alternative hypotheses take the form

$H_0 : \pi_A - \pi_B \leq -d$,

$H_1 : \pi_A - \pi_B > -d$,

which on the NNT scale equate to

$H_0 : NNT \geq -1/d$,

$H_1 : NNT < -1/d$.

In order to conclude non-inferiority, one needs to reject the null hypothesis. In practice, this the same as constructing a (1-2α)100% confidence interval and concluding non-inferiority provided that the lower end of this confidence interval is above –d on the proportional scale or greater than –1/d on the NNT.

For a simple non-inferiority trial this may seem to be operationally straightforward. However, as discussed in Chapter 1, such trials are seldom just designed as non-inferiority trials but as "as good as or better" trials where an appropriate closed testing procedure is applied to investigate both superiority and non-inferiority whilst maintaining the overall Type I error rate [Morikawa and Yohsida, 1995; CPMP, 2000]. As a consequence "as good as or better" trials also incorporate the null and alternative hypotheses of superiority trials and hence issues with using NNT for superiority trials arise.

3.4.1.3. Equivalence Trials

The null and alternative hypotheses for an equivalence trial on the proportional scale may take the form

$H_0 : \pi_A - \pi_B \geq d$ or $\pi_A - \pi_B \leq -d$,

$H_1 : -d < \pi_A - \pi_B < d$.

On the NNT scale the equivalent Null and alternative hypotheses become

$H_0 : NNT \leq 1/d$ or $NNT \geq -1/d$,

$H_1 : -1/d < NNT > 1/d$,

which operationally would equate to constructing a confidence interval and demonstrating that it fall completely within the interval ($-\infty$ to $-1/d$) and ($1/d$ to ∞)? Not what one could argue to be the most easily understood concept.

3.4.2. Confidence Intervals for Number Needed to Treat

This sub-section will discuss the different methods of calculating a confidence interval for the NNT. It will start with the standard methodology for calculating confidence intervals for the NNT of taking the reciprocal of the difference in proportions and then move onto two other methods the delta method and bootstrapping.

3.4.2.1. Reciprocal of the Confidence Intervals of the Difference in Proportions

Given that NNT is the reciprocal of $p_A - p_B$ the convention is to simply take the reciprocal of the confidence interval of $p_A - p_B$ to obtain its confidence interval [Altman, 1998].

Table 3-7. Table of confidence intervals for the difference in responses (p_A=0.4 and p_B=0.20) and number needed to treat by three methods for different sample sizes

	Confidence Intervals			
Sample Size	Proportion	Reciprocal Method	Delta Method	Bootstrapping
10	-0.19 to 0.59	-5.26 to 1.69	-4.75 to 14.75	-5.00 to 1.67
25	-0.05 to 0.45	-20.00 to 2.22	-1.25 to 11.25	-24.79 to 2.27
50	0.02 to 0.38	2.63 to 50.00	0.50 to 9.50	2.63 to 49.42
100	0.08 to 0.32	3.13 to 12.50	2.00 to 8.00	3.12 to 12.57
500	0.14 to 0.26	3.85 to 7.14	3.50 to 6.50	3.91 to 7.02
1000	0.16 to 0.24	4.17 to 6.25	4.00 to 6.00	4.19 to 6.25

The problem with this approach is that at first glance confidence intervals are obtained that do not contain the point estimate for negative studies. The first two lines of Table 3.7 illustrate this (column 3). For example for a NNT point estimate of 5 a confidence interval of -5.26 to 1.69 is obtained. This problem with negative results has led to confidence intervals sometimes just being just quoted for positive (statistically significant) trials. The interpretation of negative results is in fact in terms of "to infinity and beyond" with confidence intervals for the given example actually being ($-\infty$ to -5.3] and [1.7 to ∞) [Altman, 1998] and is due to the fact that the "null" value of a $NNT = \infty$ (and negative trials should contain the null). The two part confidence

interval does now actually include the point estimate - although interpretation may be a bit difficult.

3.4.2.2. The Delta Method

To first reconsider the calculation of the confidence interval for the NNT one should first consider how to obtain a variance estimate. For a given function $f(x)$, where $x = p_A - p_B$ and $f(x) = NNT$, the variance for the number needed to treat can be defined as [Armitage and Berry, 1987]

$$\text{var}(NNT) = \text{var}(1/x) = \text{var}(1/(p_A - p_B)) \approx (f'(x))^2_{x=E(x)} \text{var}(x),$$

giving [Lesaffre and Pledger, 1999; Schulzer and Mancini, 1996]

$$\text{var}(NNT) = \text{var}(p_A - p_B)/(p_A - p_B)^4 = (NNT)^4 \text{var}(p_A - p_B). \qquad (3.3.2)$$

Thus, the standard statistic can be defined as

$$\frac{NNT}{se(NNT)} = \frac{1}{p_A - p_B} \frac{(p_A - p_B)^2}{se(p_A - p_B)} = \frac{p_A - p_B}{se(p_A - p_B)}. \qquad (3.3.3)$$

Which is identical to the standard statistic for the difference in proportions. One can thus calculate the confidence interval for the number needed to treat from,

$$\text{NNT+/-}\, Z_{1-\alpha/2}\, \text{se(NNT)} = NNT \pm Z_{1-\alpha/2} (NNT)^2 se(p_A - p_B),$$

which is the same as calculating the CI for the difference in proportions and multiplying it by $(NNT)^2$. This approach is sometimes referred to as the delta method [Matthews, 2000]. It is just as easy as taking reciprocals to calculate and on the face of it gives more intuitive answers as the confidence interval always includes the point estimate. The equivalent confidence interval to the example quoted earlier is –4.8 to 14.8. In fact this confidence interval still has the "to infinity and beyond" interpretation in that it should actually be (-∞to –4.8] and (5 to 14.8].

One issue with the Delta method is that as Lesaffre and Pledger [1999] point out, for any sample size and any expected proportional difference there is always a non-zero probability that the observed proportional difference will equal zero. Hence, for any expected proportional difference and sample size the expected value of NNT is infinity along with its variance.

3.4.2.3. Bootstrapping

A final approach for calculating confidence intervals is to calculate them using non-parametric bootstrap methodology – extending the work from other areas [Efron and

Tibshirani, 1993; Keene, 2002; Julious, 2001]. To calculate a bootstrap confidence interval a bootstrapping sample, with replacement, is taken from the data. This bootstrap sample is of the same sample size as the original data. For each bootstrap sample a NNT is calculated. This is repeated a large number of times, 10,000 say, to generate a bootstrap distribution for NNT.

The bootstrap distribution of the NNTs is then ordered. As on the NNT scale 1>2>3>... etc the data are "reverse" ordered separately for positive and negative simulated values. For example for 5 NNTs of 2,-1,-4,3 and 8 the data would be ordered as –1,-4,8,3 and 2. After this ordering the percentiles were taken. To calculate bootstrap 95% confidence interval the values associated with the 2.5% to the 97.5% percentiles are taken. This bootstrap method is known as the percentile method [Efron and Tibshirani, 1993].

It may be worth noting that the apparently unusual ordering of NNT does occur on other scales. For example on the log scale 0.25>0.5>0.75 for observed effects which one counters by log transforming to get the data on an arithmetic scale. For NNT one would reciprocate the data (i.e. back to the $p_A - p_B$ scale) to get a conventional arithmetic ordering.

When calculating a bootstrap confidence interval, bootstrap samples from a finite sample size with only two categories of outcome are taken. Thus, a number of ties will be expected. As bootstrapping requires the calculation of the number of NNTs below a certain value ties can represent a problem in calculations. To overcome this problem a random number is added to each bootstrap $p_A - p_B$ response prior to the calculation of the bootstrap NNT. In the examples given here a random number sampled from the uniform distribution between (-0.0005, 0.0005) was added to each $p_A - p_B$.

Bootstrapping has been commented on as a way of calculating a confidence interval for NNT by Hutton [2000]. However, Hutton also highlighted the issue of indeterminable NNTs for bootstrap samples where $p_A - p_B = 0$. Through adding a small random number one overcomes this problem through in effect randomly assigning bootstrap samples where $p_A - p_B = 0$ to either very large positive or negative NNTs.

The equivalent bootstrap confidence interval to the example quoted earlier for methods 1 and 2 is –5.00 to 1.67. The estimated bootstrap confidence interval for this example is quite close to method 1. This bootstrap calculation was undertaken in the Interactive Matrix Language in SAS [1985].

3.4.2.4. Comparison of the Different Methods

A simple comparison of the three methods of calculating the confidence intervals is given in Table 3.7. From this table it is evident that one feature of the delta method is that the confidence interval is symmetric about the point estimate. Simple reciprocation of the proportional difference confidence intervals and bootstrapping give similar confidence interval estimates, which differ from the delta method when the sample size

is small but all three methods start to converge as the sample size gets larger and the evidence towards a positive effect increases.

To further compare the three methods of calculating a confidence interval around NNT simulations were undertaken in the Interactive Matrix Language in SAS [1985] to compare with Table 3.7. As the number needed to treat is an abstract parameter on an abstract scale the simulations were done on the proportional scale with a NNT calculated for each simulation from the proportional difference. The data were simulated the respective treatment responses for each simulation taken as p_A=0.40 and p_B=0.20.

One issue with the simulations on the binary scale is that one encounters zero for the treatment differences which causes a problem when reciprocating to calculated NNT. For the bootstrapping described earlier this problem was overcome through adding a very small value to all simulated proportional differences. For this simulation exercise the problem was overcome by randomly assigning $p_A - p_B = 0$ to be either a very large negative or very large value for NNT. For a given sample size the simulation was repeated 10,000 times.

After the simulation for each sample size the data were ordered smallest to largest and five percentiles (2.5th, 25th, 50th, 75th and 97.5th) were taken. The 2.5th and 97.5th percentile should correspond to the 95% confidence interval. The simulated NNT were ordered in the same why as that described for bootstrap samples. The results are given in Table 3.8.

Table 3-8. Results from simulations for number needed to treat

	Percentile				
Sample Size	2.5	25	50	75	97.5
10	-5.00	10.00	5.00	3.33	1.67
25	-25.00	8.33	5.00	3.57	2.27
50	50.00	7.14	5.00	3.85	2.63
100	12.50	6.25	5.00	4.17	3.03
500	6.94	5.49	5.00	4.55	3.90
1000	6.21	5.35	5.00	4.67	4.17

For negative trials the simulations are awry from method 2 but for positive and larger trials there is greater agreement. The issue with the simulations though is that each NNT was estimated indirectly from the proportional difference, as it would be done practically, whilst method 2 provides a confidence interval for NNT assuming that in future trials NNT could be directly derived.

3.4.2.5. *Further Issues with Number Needed to Treat*

One of the problems with NNT is the fact that it is just one number, which is sometimes quoted as an advantage. Usually when one quantifies a treatment effect it is usually in context with something. For example in a hypertension trial if subjects on treatment A had a mean diastolic blood pressure of 110 whilst subjects on treatment B had a mean of 100, a treatment effect of 10 would be observed. In this instance the number 10 has a context in terms of the effects observed on the individual treatments. However, what is the context of NNT=10? Table 3.9 summarises this point. For different values of p_A and p_B equivalent values of NNT are derived. As the NNTs are the same one would imagine that for each, the effects would be the same. However, an effect of 0.10 in terms of $p_A - p_B$ ranges from a trebling in the proportion observed to an increase of just a third when putting the effect in context with p_A and p_B. A marked difference in interpretation and meaning. To assist in interpretation therefore NNT should not be quoted in isolation but in context with p_A, p_B and $p_A - p_B$ [Julious, 2005c] indeed due to the issues raised in this Chapter in isolation a NNT should really stand for Nonsensical Numeration of Treatment.

Table 3-9. Proportional differences that would give an equivalent estimate for the number needed to treat

p_A	p_B	$p_A - p_B$	NNT
0.05	0.15	0.10	10
0.10	0.20	0.10	10
0.20	0.30	0.10	10
0.30	0.40	0.10	10

3.5. Odds-Ratio

The difference between two proportions can also be expressed through the odds ratio (OR) which is defined as

$$OR = \frac{p_B(1-p_A)}{p_A(1-p_B)}. \qquad (3.4.1)$$

Odds of 2:1 would mean that for every 3 subjects on a control regimen, for example, one would expect one event i.e. non-events are twice as numerous as events. Odds of 4:1 on the investigative regimen would mean non-events are four times as numerous. An odds-ratio is simply therefore a ratio of two odds and is an assessment of the likelihood of success on one treatment compared to another. Hence, an odds-ratio would be 2 indicating non-events relative to events are twice as numerous on the investigative regimen compared to control.

One of the main advantages of the OR is that it is invariant to the definition of success [Olkin, 1998; Walker, 1998]. An analysis based on the OR also easily allows one to

adjust for covariates such that estimates can be provided that are independent of, but adjusted for, any predictive factors of interest. As an analysis with covariate adjustment is often the standard analysis it supports the idea of the OR being the standard parameter.

The log-odds-ratio is also an attractive scale for analysis as it is both unbounded and likely to be additive across a wide range.

3.5.1. Odds-Ratios and Clinical Trials

This sub-section will discuss how a study, in terms of the null and alternative hypothesis, would be designed where the odds-ratio is the primary assessment of efficacy.

3.5.1.1. Superiority Trials

When making inferences with odds-ratios it is much simpler to work on the Log(OR) scale then back transform to the original scale. Hence, when working on the odds ratio scale the null hypothesis and alternative hypotheses can be defined as

$H_0 : Log(OR) = 0$,

$H_1 : \text{Log(OR)} = \text{d}$,

where d is some pre-defined treatment effect used for sample size calculations.

3.5.1.2. Non-Inferiority Trials

For a non-inferiority trial the null and alternative hypothesis would be expressed as the following

$H_0 : Log(OR) \leq -d$,

$H_1 : Log(OR) > -d$,

where d is a pre-define non-inferiority limit.

3.5.1.3. Choice of Non-Inferiority Limit

It is when considering non-inferiority trials (and equivalence trials later) that one sees one an advantage of working on the odds-ratio scale. As discussed earlier in the chapter there are hard regulatory guidance for the antimicrobial therapeutic area to assess non-

inferiority as given in Table 3.2. The guidance from the FDA [1992] is currently a function dependent on the anticipated control response rate.

This step function does present problems when working with absolute differences. Suppose one designed a trial based on an anticipated active response rate of 78% and a margin of 20% but one actually observed 82%. This brings one over into the next margin level of 15%. Also, the step function states one can use the same margin whether the anticipated response rate is 80% or 89%. One could argue these are markedly different response rates.

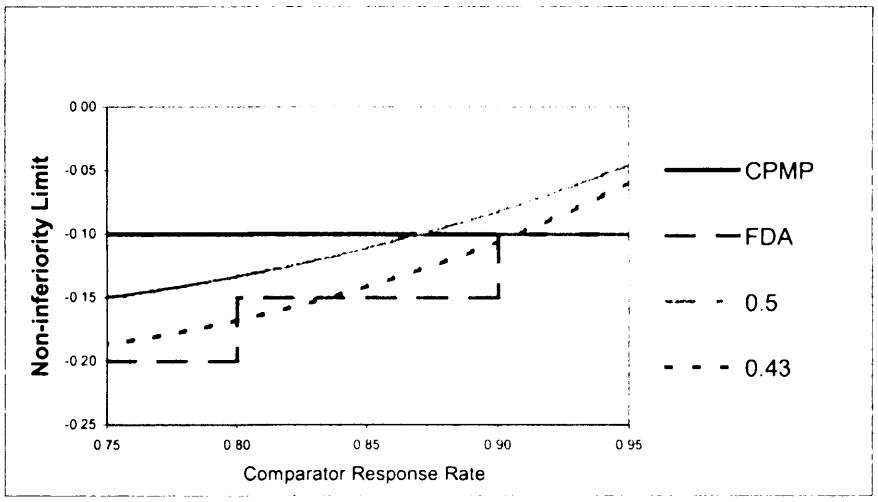
Working on the odds-ratio scale avoids the problems with the issues with stepped non-inferiority margin. This is because on the odds-ratio scale a fixed margin would equate to different margins on the proportional scale. This has been recognised by a number of authors. Garrett [2003] has recommended using a margin of 0.5 on the odds-ratio scale whilst Senn [1997] has recommend a margin of 0.55 and Tu [1998] a margin of 0.43. The relative merits of these margins can be seen in Table 3.10 and Figure 3.7.

Table 3-10. Table of differences on the proportional scale that are equivalent to different odds-ratios for various anticipated expected responses on one treatment arm

	Odds-Ratio				
p_A	0.40	**0.45**	**0.50**	**0.55**	0.60
0.95	0.066	**0.054**	**0.045**	**0.037**	0.031
0.90	0.117	**0.098**	**0.082**	**0.068**	0.056
0.85	0.156	**0.132**	**0.111**	**0.093**	0.077
0.80	0.185	**0.157**	**0.133**	**0.113**	0.094
0.75	0.205	**0.176**	**0.150**	**0.127**	0.107
0.70	0.217	**0.188**	**0.162**	**0.138**	0.117
0.65	0.224	**0.195**	**0.169**	**0.145**	0.123
0.60	0.225	**0.197**	**0.171**	**0.148**	0.126
0.55	0.222	**0.195**	**0.171**	**0.148**	0.127
0.50	0.214	**0.190**	**0.167**	**0.145**	0.125

Table 3.10 gives the equivalent difference on the proportional scale for different odds-ratios and control response rates. Taking the 0.45 column to approximately represent the Tu margin one can see the relative merits of each margin particularly at the different step points, given in Table 3.2, from the FDA guidance [1992]. The Tu [1998] margin gives the closest agreement at the step points to the FDA guidance – although it equates to an 11% margin at the 0.90 step and 17% at 0.80 which are too high. The margin of Senn is the most conservative and would guarantee that the difference is no greater than 15% no matter what the control prevalence is. Finally, the margin of Garrett falls between those of Senn and Tu although has the advantage of being a round number. Figure 3.7 figuratively demonstrates these points being a repeat of Figure 3.1 although with the margins of Garrett [2003] and Tu [1998] now included.

Figure 3-7. Graphic illustration of CPMP and FDA non-inferiority limits on the proportional scale for fixed odds-ratios



Incidentally a margin which has never been mentioned is one of 0.47 which gives margins that are closest to all the steps without crossing them: for 70% it would equate to 17.7%; 80% and 14.7% and 90% and 9.1%.

3.5.1.4. Equivalence Trials

Similarly to the proportional scale the null and alternative hypotheses for an equivalence trial may take the form

$H_0 : \log(OR) \geq d$ or $\log(\mathrm{OR}) \leq -d$,

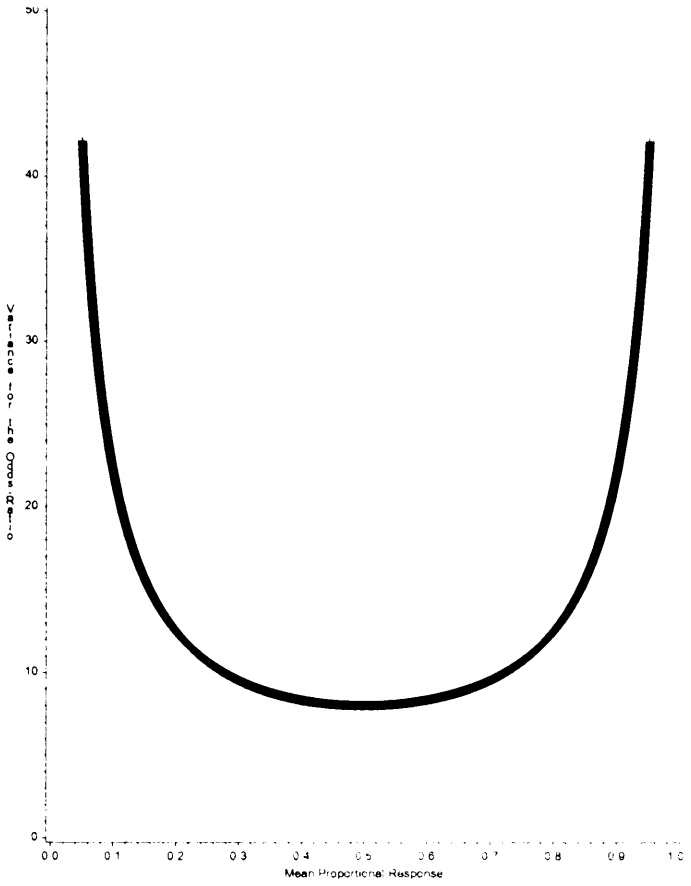
$H_1 : -d < \log(OR) < d$,

with margins set similarly to those of non-inferiority trials described previously.

3.5.2. Calculation of Confidence Intervals

Methods for the calculation of confidence intervals for the odds-ratio will be described. Two methods will be discussed in detail. That of Normal approximation (on the log scale) and the exact methodology.

Figure 3-8. Plot of the variance of a log odds ratio for different mean proportional responses



3.5.2.1. Normal Approximation

Under the Normal approximation the confidence interval for the log(OR) is defined as

$$\log(OR) \pm Z_{1-\alpha/2} se(\log(OR)), \qquad (3.4.2)$$

where the confidence interval on the original odds-ratio scale is obtained by back transforming the confidence interval on the log scale. There are a number of ways of estimating the standard error for the odds-ratio [McCullagh, 1980] although in this chapter will concentrated on just one, defined by Whitehead [1993] as

$$\text{var}(LogOR) = \frac{12}{n\left(1 - \sum_{i=1}^{2} \bar{p}_i^3\right)}, \qquad (3.4.3)$$

where $\bar{p}_1 = (p_A + p_B)/2$ and $\bar{p}_2 = 1 - \bar{p}_1$. As will discussed in chapter 5 this formula can be generalised to more that a 2 point response variable. For just two categories (3.4.3) becomes

$$\frac{12}{n\left(1-\sum_{i=1}^{2}\bar{p}_i^3\right)} = \frac{12}{n\left((\bar{p}_1+\bar{p}_2)^3-\bar{p}_1^3-\bar{p}_2^3\right)} = \frac{4}{n\bar{p}_1(1-\bar{p}_1)}$$

where $2\bar{p}_1(1-\bar{p}_1) \approx p_A(1-p_A)+p_B(1-p_B)$. Thus, the variance for the log(OR) is proportional to the reciprocal of the variance for the difference in proportions i.e. $\text{var}(LogOR) \propto 1/\text{var}(p_A - p_B)$.

This relationship is highlighted in Figure 3.8 which gives the variance for log(OR) for different mean proportional responses. As one can see from this figure, the variances take a U shape in relation to the mean proportional response with a minimum when $\bar{p} = 0.5$. This is the opposite to the proportional scale where the variance takes the maximum at the mid point.

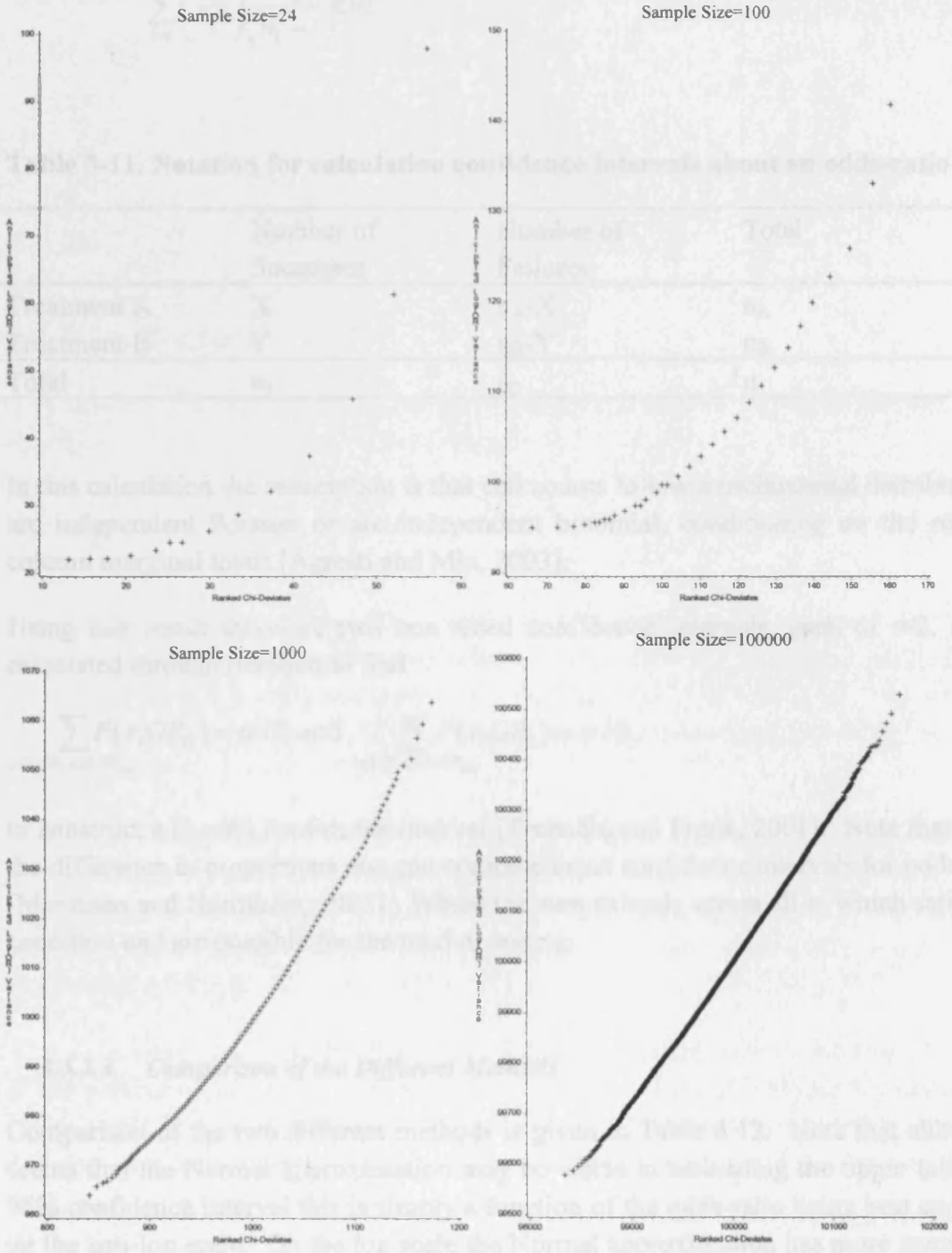
Another feature to highlight is how the variance is relatively stable for a large of the range of the mean response (0.3 to 0.7). Only varying greatly as the mean response approaches a boundary.

An empirical investigation of the asymptotic assumptions about the variance is given in Figure 3.9. The simulation was undertaken in SAS [1990]. For each sample size 10,000 simulations were undertaken assuming the response rate to be 60% and for each simulation the following was calculated

$$(n-1)\frac{4}{n\bar{p}_1(1-\bar{p}_1)}\Big/\frac{4}{n\bar{\pi}_1(1-\bar{\pi}_1)} = (n-1)\frac{\bar{\pi}(1-\bar{\pi})}{\bar{p}(1-\bar{p})},$$

where $\bar{p}$ is the estimated mean response from the simulation and $\bar{\pi}$ is the population mean prevalence ($\bar{\pi}$ =0.60) from which each simulation was drawn. What is evident from this figure is that the approximation to the chi-distribution is quite weak for small sample sizes but improves the larger the trial. The empirical investigation suggest that the approximation to the chi-squared is slightly better for the variance of the log odds-ratio compared to the proportional difference discussed earlier.

Figure 3-9. Chi-probability plots for different sample sizes for a variance for the log(OR) for different mean responses (π=0.6) sampled from a binomial distribution



3.5.2.2. Exact Confidence Intervals

Following the notation in Table 3.11, by conditioning on x a sufficient statistic for the odds-ratio can be obtained [Fisher, 1935; Chan, 2003; Dunnett and Gent, 1977]. Hence, the probability of observing an outcome in the top left cell equal to x can be calculated from an hypergeometric distribution [Toendle and Frank, 2001; Agresti and Min, 2003; Agresti, 2001]

$$P(x;OR) = \frac{\binom{n_A}{x}\binom{n_B}{n_1 - x}OR^x}{\sum_{i=0}^{n_1}\binom{n_A}{i}\binom{n_B}{n_1 - i}OR^i}. \qquad (3.4.4)$$

Table 3-11. Notation for calculation confidence intervals about an odds-ratio

	Number of Successes	Number of Failures	Total
Treatment A	X	n_A-X	n_A
Treatment B	Y	n_B-Y	n_B
Total	n_1	n_2	n

In this calculation the assumption is that cell counts follow a multinomial distribution or are independent Poisson or are independent binomial, conditioning on the row and column marginal totals [Agresti and Min, 2003].

Using this result therefore two one sided confidence intervals, each of $\alpha/2$, can be calculated through iteration to find

$$\sum_{(x):OR(x)\le OR_{obs}} P(x;OR_U) = \alpha/2 \text{ and } \sum_{(x):OR(x)\ge OR_{obs}} P(x;OR_L) = \alpha/2,$$

to construct a (1-α)% confidence interval [Troendle and Frank, 2001]. Note that unlike the difference in proportions one can compute exact confidence intervals for odds-ratios [Miettinen and Nurminen, 1985]. Where the sum extends across all x, which satisfy the condition and are possible for the total n_1 and n_A.

3.5.2.3. Comparison of the Different Methods

Comparison of the two different methods is given in Table 3.12. Note that although it seems that the Normal approximation may be worse at estimating the upper tail of the 95% confidence interval this is simply a function of the odds-ratio being best expressed on the anti-log scale. On the log scale the Normal approximation has more comparable precision for both the upper and lower tail compared to the exact methodology.

As with previous cases in this chapter the Normal approximation gets close to that of the exact confidence intervals as the sample size gets larger although for all sample sizes it does give a narrower range for the limits. After a sample size per group of around 50 the Normal approximation and the exact method are quite close.

Table 3-12. Table of confidence intervals for different proportions (p_A=0.50 and p_B=0.33) equating to an odds-ratio of 2 by two methods for different sample sizes per group

	Confidence Intervals	
Sample Size	Normal Approximation	Exact
12	0.39 to 10.14	0.29 to 14.61
24	0.63 to 6.30	0.54 to 7.61
48	0.89 to 4.50	0.81 to 4.97
96	1.13 to 3.55	1.07 to 3.74
192	1.33 to 3.00	1.30 to 3.09
384	1.50 to 2.66	1.48 to 2.71
768	1.63 to 2.45	1.62 to 2.47

3.6. Relative Risk

The relative difference between two proportions can also be expressed through simply taking their ratio, termed the relative risk (RR)

$$RR = \frac{p_A}{p_B}. \qquad (3.5.1).$$

One of the main advantages of the relative risk is that it is allegedly easier to interpret. However, in absolute terms the same relative risk could be quite markedly different with a relative risk of 2 equally be ratios of 0.002/0.001, 0.2/0.1 or 0.80/0.40. This alone can cause problems in terms of interpretation (especially rare events) and design as will be discussed later.

It is worth noting here the fact one of the main criticisms made of the odds-ratio (OR), , with reference to its interpretation is that it is not a relative-risk [Davies, Crombie and Tavaloli, 1998; Sackett, Deeks and Altman, 1996; Altman, Deeks and Sackett, 1998; Deeks, 1998]. Although the fact that an OR does not equal a RR should not be considered a criticism.

The main disadvantage of the RR is that it is not invariant to the definition of success. If the number of events on two respective treatments were 60/120 and 80/120 then the relative risk for this would be 0.75. However, if it is the number of none events that is of importance then the relative risk becomes 1.33 (the odds-ratio is 2 no matter what the choice of success is).

3.6.1. Relative Risk and Clinical Trials

The issues with using the relative risk in clinical trials will now be highlighted. As with odds-ratios, it is the log-relative-risk which is the most attractive scale, to base inference.

3.6.1.1. Superiority Trials

When one is thinking in terms of the relative risk, or as written here log relative risk, the null and alternative hypothesis should be express as

$H_0 : \log(RR) = 0,$

$H_0 : \log(RR) = d,$

where d is some pre-define treatment effect.

3.6.1.2. Non-Inferiority Trials

For a non-inferiority trial on the relative scale the null and alternative hypotheses take the form

$H_0 : \log(RR) \leq -d,$

$H_1 : \log(RR) > -d,$

where d is some predefined non-inferiority limit.

3.6.1.3. Choice of Non-Inferiority Limit

It is in the assessment of non-inferiority trials that the rationale for using the relative risk begins to fail. Remember that for an odds-ratio one could set a fixed non-inferiority limit, which would have a consistent meaning regardless of the control response rate. In particular Table 3.11 demonstrates that an odds-ratio of 0.5, say, for a non-inferiority limit would allow one to meet all the steps in the FDA non-inferiority guidance given in Table 3.2.

The relative risk does not share the properties of the odds-ratio. For example a relative risk of 2.0 would equate to a 35% difference on the proportion scale if active response rate was 70% but a 40% difference if the active response rate was 80%. Thus, the relative risk non-inferiority boundary would need to change for different active response rates. Bringing in the need for a step function for non-inferiority limits. This feature alone negates the use of the relative risk as a statistic to use in clinical trials.

3.6.1.4. Equivalence Trials

The null and alternative hypotheses for an equivalence trial on the relative scale may take the form

$H_0 : \log(RR) \geq d \text{ or } \log(RR) \leq -d,$

$H_1 : -d < \log(RR) < d$,

where d is some pre-defined equivalence limit. Equivalence trials however, share the same issues as non-inferiority trials in that the equivalence margin will need to change depending on the control response rate.

3.6.2. Calculation of Confidence Intervals

The most common approach is to calculate the confidence interval, under Normal approximation, on the log scale using

$$\log(RR) \pm Z_{1-\alpha \cdot 2} se(\log(RR)), \qquad (3.5.2)$$

and then back transforming back to get a confidence interval on the original scale. The standard error for the logged relative risk can be taken as

$$\text{where } se(\log(RR)) = \sqrt{\frac{1}{E_A} + \frac{1}{E_B}}, \qquad (3.5.3)$$

where E_A and E_B are the observed number of events on treatments A and B respectively. The standard error is estimated using the delta method described earlier in this chapter. For alternative methods for calculating confidence intervals please see the work of Graham, Mengersen and Morton [2003] and Ederer and Mantel [1974].

3.7. Summary of Chapter 3

This chapter described four ways of summarising binary data:

1. Absolute risk reduction
2. Odds-ratio
3. Relative risk reduction
4. Number needed to treat

Due to the relative merits of each summary measure in context with the different types that may be conducted it is recommended that relative risk and the number needed to treat should not be used to summarise individual trials. Instead the summary measures of absolute risk and odds-ratio should be used. The odds-ratio from a statistical perspective is particularly attractive as it allows for a transition through different non-inferiority (and equivalence) criteria for a given active control response rate. Subsequent chapters will only now discuss the odds-ratio and absolute risk reduction.

$H_1 : -d < \log(RR) < d$,

where d is some pre-defined equivalence limit. Equivalence trials however, share the same issues as non-inferiority trials in that the equivalence margin will need to change depending on the control response rate.

3.6.2. Calculation of Confidence Intervals

The most common approach is to calculate the confidence interval, under Normal approximation, on the log scale using

$$\log(RR) \pm Z_{1-\alpha/2} se(\log(RR)), \qquad (3.5.2)$$

and then back transforming back to get a confidence interval on the original scale. The standard error for the logged relative risk can be taken as

$$\text{where } se(\log(RR)) = \sqrt{\frac{1}{E_A} + \frac{1}{E_B}}, \qquad (3.5.3)$$

where E_A and E_B are the observed number of events on treatments A and B respectively. The standard error is estimated using the delta method described earlier in this chapter. For alternative methods for calculating confidence intervals please see the work of Graham, Mengersen and Morton [2003] and Ederer and Mantel [1974].

3.7. Summary of Chapter 3

This chapter described four ways of summarising binary data:

1. Absolute risk reduction
2. Odds-ratio
3. Relative risk reduction
4. Number needed to treat

Due to the relative merits of each summary measure in context with the different types that may be conducted it is recommended that relative risk and the number needed to treat should not be used to summarise individual trials. Instead the summary measures of absolute risk and odds-ratio should be used. The odds-ratio from a statistical perspective is particularly attractive as it allows for a transition through different non-inferiority (and equivalence) criteria for a given active control response rate. Subsequent chapters will only now discuss the odds-ratio and absolute risk reduction.

This chapter also discussed the relative asymptotic properties of data, which take a binomial form. It highlighted how for large sample sizes binary data have asymptotic properties, which from a practical point of view, means that the approximate methodologies agree with the exact methodologies. A consequence of this is that as most clinical trials are relatively large the sample size methodologies assuming asymptotic results will hold.

If exact confidence intervals are to be calculated for a single response then this chapter recommends that a Beta distribution be used. Operationally these are relatively straightforward to calculate and the calculations can be undertaken in most statistical packages.

The issue of the asymptotic properties for a binary response estimated from a relatively small sample size becomes a concern if retrospective trial data (from a small trial) are to be used to design a prospective study. In Chapter 2 it was highlighted how a result using the non-central t distribution could be used if a sample variance was to be used in the calculations. Here, the data could be assumed to take a Normal form. Part of the proof of this result was the fact the ratio of the sample variance over the population variance could be assumed to take a chi-squared distribution. In this chapter it was highlighted that the sample variance does not follow a chi-squared distribution for binary data with small sample sizes. And hence questions whether the result for Normal data discussed in Chapter 2 could be extended to binary data. This will be discussed in detail in Chapter 4.

4. CHAPTER 4 - SAMPLE SIZE CALCULATIONS FOR CLINICAL TRIALS WITH BINARY DATA

This chapter describes the sample size calculations for trials where the primary outcome is binary. As discussed in Chapter 3 binary outcomes are common endpoints in clinical trials and appear when the outcome of interest is a two point response variable such as presence/absence, alive/dead or yes/no.

The sample size calculations will be described for situations where the trial design is either parallel group or a cross-over. The conventional calculations will first be introduced for each trial design and objective, followed by calculations, which account for the imprecision of the estimates.

4.1. Aims of the Chapter

The main issues to be covered in this chapter are:

- A description of standard sample size calculations for clinical trials with a binary primary outcome.
- A review of methodologies for calculation of sample sizes for cross-over trials with an assessment of how effect sizes for parallel group studies can be generalised to cross-over trials.
- An assessment of how the effect of period can impact the design and analysis of a cross-over study.
- An investigation (and make recommendations for) the calculation of sample sizes for non-inferiority and equivalence trials.
- An assessment of whether methodologies for imprecisely estimated variances developed for data anticipated to take a Normal form in Chapter 2 can be extended binary data.
- To investigate methodologies to assess the sensitivity of studies to the assumptions made in the sample size calculations.
- To develop numerical methods for the estimation of sample sizes that account for the imprecision in the control response.
- To develop Bayesian methodologies to assess a study's sensitivity and also to account for the imprecision in the estimate of the control response.
- To investigate the effect of factors such as covariates in the design and analysis of studies with a binary response.

4.2. Superiority Trials

As discussed in Chapter 3 there are a number of ways to summarise binary data for a parallel group trial, such as odds-ratios, relative risks, the number needed to treat and the absolute difference in the proportions. The two summary statistics that this chapter will concentrate on are those of the odds-ratio and the difference in the proportions.

4.2.1. Parallel Group Trials

4.2.1.1. Sample Sizes with the Population Effects Assumed Known

4.2.1.2. Odds-ratio

For a given binary response, p_A and p_B are defined as the proportion of responders expected on each of the two treatment groups, A and B, respectively. Each of these expected responses can, in turn, be written in terms of odds, $p_A/(1-p_A)$ and $p_B/(1-p_B)$ (each odds is a ratio of responders over non-responders). As a consequence an odds-ratio (OR) can be used as assessment of the treatment difference (and effect size for sample size calculations) - where the OR is defined as

$$OR = \frac{p_A(1-p_B)}{p_B(1-p_A)}.$$

In a trial where the objective is to determine whether there is evidence of a statistical difference between the regimens the null (H_0) and alternative (H_1) hypotheses may take the form:

H_0: The two treatments have equal effect with respect to the odds-ratio ($OR = 1$).

H_1: The two treatments are different with respect to the odds-ratio ($OR \neq 1$).

From these null and alternative hypotheses a formula can be constructed to calculate a sample size per group [Julious, George, Machin et al, 1997; Julious, Walker, Campbell et al, 2000; Campbell, Julious and Altman, 1995; Machin, Campbell, Fayers et al, 1997; Whitehead, 1993]. As discussed in Chapter 1, in general terms for a 2-tailed, α -level test the variance of the measure of effect must satisfy

$$\text{Var(S)} = \frac{d^2}{(Z_{1-\beta} + Z_{1-\alpha/2})^2}, \qquad (4.1.1)$$

and the variance of the log-odds-ratio (S=log-odds-ratio in this instance) can be approximated by [Whitehead, 1993]

$$Var(S) = \frac{6}{n\left(1 - \sum_{i=1}^{2} \bar{p}_i^3\right)}, \qquad (4.1.2)$$

where $\bar{p}_i$ is the average response across treatments for each outcome category (i.e. $\bar{p}_1 = (p_{1_A} + p_{2_B})/2$ and $\bar{p}_2 = 1 - \bar{p}_1$) and α and β are the overall type I and type II errors respectively with $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ denoting the percentage points of a standard Normal distribution for these two errors. Here n is the sample size per group. Note that in this chapter the assumption will be that there is equal allocation to treatment.

Now by equating (4.1.1) with (4.1.2) one requires

$$n = \frac{6\left[Z_{1-\beta} + Z_{1-\alpha/2}\right]^2 / (\log OR)^2}{\left[1 - \sum_{i=1}^{2} \bar{p}_i^3\right]} \qquad (4.1.3)$$

4.2.1.3. Proportional Difference

Keeping the data on the proportional scale the treatment difference would be expressed in terms of an absolute difference in the proportions defined as $p_A - p_B$. On this scale the null and alternative hypothesis would be written as:

H_0: The two treatments have equal effect with respect to the proportional response ($\pi_A = \pi_B$).

H_1: The two treatments are different with respect to the proportional response ($\pi_A \neq \pi_B$).

Table 4-1. Sample size estimates for one arm of a parallel group trial for various expected outcome responses for a given treatment (p_A) and odds-ratios for a two sided type I error rate of 5% and 90% power

	Odds-Ratio					
p_A	1.25	1.50	1.75	2.00	3.00	4.00
0.10	5112	1653	914	621	276	184
0.20	2819	900	494	333	146	98
0.30	2105	663	360	242	105	69
0.40	1803	560	300	200	85	56
0.50	1694	517	274	180	75	49
0.60	1725	517	270	176	70	45
0.70	1926	566	290	186	71	44
0.80	2468	709	356	224	81	49
0.90	4278	1199	588	362	121	68

From these a hypotheses a sample size formula - following the same arguments as (4.1.1) and (4.1.2) - where an absolute difference is the response of interest can be derived [Julious, George, Machin et al, 1997; Campbell, Julious and Altman, 1995; Machin, Campbell, Fayers et al, 1997]

$$n = \frac{\left[Z_{1-\beta} + Z_{1-\alpha/2}\right]^2 \left(p_A(1-p_A) + p_B(1-p_B)\right)}{(p_A - p_B)^2}. \qquad (4.1.4)$$

Table 4.1 gives a table of sample sizes for various odds-ratios and responses on a given treatment using (4.1.3). Table 4.2 gives the sample sizes using (4.1.4).

Table 4-2. Sample size estimates for one arm of a parallel group trial for various expected outcome responses for a given treatment (p_A) and comparator (p_B) for a two sided type I error rate of 5% and 90% power

	p_B								
p_A	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.10	578								
0.15	184	915							
0.20	97	263	1209						
0.25	63	120	331	1461					
0.30	44	79	158	389	1671				
0.35	33	54	94	182	437	1839			
0.40	25	39	62	106	200	473	1965		
0.45	20	29	44	69	115	214	500	2048	
0.50	16	23	33	48	74	121	223	515	2091

4.2.1.4. Equating Odds-Ratios with Proportions

Although (4.1.3) and (4.1.4) seem on the face of it to be quite different it can be shown that they are approximately algebraically the same [Julious and Campbell, 1996]. This comes from the following two results

$$\frac{6}{\left(1 - \sum_{i=1}^{2} \bar{p}_i^3\right)} = \frac{6}{(\bar{p}_1 + \bar{p}_2)^3 - \bar{p}_1^3 - \bar{p}_2^3} = \frac{2}{\bar{p}_1(1-\bar{p}_1)} = \frac{2}{\bar{p}_1(1-\bar{p}_1)}, \qquad (4.1.5)$$

and

$$\log(OR) \approx \frac{2(OR-1)}{OR+1},$$

which holds for $0.33 \leq OR \leq 3.00$ (i.e. for most practical differences when design a clinical trial). Thus,

$$\frac{2(OR-1)}{OR+1} = \frac{2(p_A - p_B)}{p_A(1-p_B) + p_A(1-p_B)} \approx \frac{p_A - p_B}{\bar{p}_1(1-\bar{p}_1)}, \qquad (4.1.6)$$

Which if substituted back into (4.1.5) gives the result

$$n = \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \left(\frac{2}{\bar{p}_1(1-\bar{p}_1)}\right)\left(\frac{\bar{p}_1(1-\bar{p}_1)}{p_A - p_B}\right)^2 = \frac{2\bar{p}_1(1-\bar{p}_1)\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)}{(p_A - p_B)^2},$$

$$\approx \frac{\left[Z_{1-\beta} + Z_{1-\alpha/2}\right]^2 (p_A(1-p_A) + p_B(1-p_B))}{(p_A - p_B)^2}.$$

Thus, (4.1.3) and (4.1.4) can be used interchangeably depending on preference. Due to this property one therefore requires

$$\frac{2\bar{p}_1(1-\bar{p}_1)\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)}{(p_A - p_B)^2} \approx \frac{2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{(\log(OR))^2 \bar{p}_1(1-\bar{p}_1)}. \qquad (4.1.7)$$

Hence, from (4.1.7) the following approximate result can be derived,

$$|p_A - p_B| \approx |\log(OR)|(\bar{p}_1(1-\bar{p}_1)), \qquad (4.1.8)$$

and therefore the proportional difference can be written in terms of the odds-ratio and the mean overall response.

As a brief note a by-product of the results highlighted in this section is that as a result of (4.1.8) the null and alternative hypotheses on the proportional scale can be written as:

H_0: The two treatments have equal effect with respect to the proportional response $|\pi_A - \pi_B| = 0$.

H_1: The two treatments are different with respect to the proportional response $|\pi_A - \pi_B| = |\log(OR)|(\bar{\pi}_1(1-\bar{\pi}_1))$.

The practical consequence of these results is that the formulae for the odds-ratio and the absolute difference can be used, for all intents and purposes, interchangeably for the same effects. This result is particularly useful for non-inferiority and equivalence trials to be discussed later in this chapter.

4.2.1.5. Worked Example

An investigator wishes to design a study where the anticipated response on the control therapy is 50%. The effect of interest is an odds-ratio of 1.5 in favour of the investigative therapy and the investigator wishes to design the study with Type I and II errors fixed at 5% and 10% respectively. From Table 4.1 one can see that the sample size required would be 517 patients per arm of the trial.

With a response rate of 50% anticipated on control an odds-ratio of 1.5 would equate to an investigative response rate of 40% or a 10% reduction. From Table 4.2 one can see that the sample size required to detect this difference is 515 patients. This sample size is approximately the same as that using the odds-ratio formula.

4.2.1.6. Sensitivity Analysis about the Estimates of the Population Effects Used in the Sample Size Calculations

In Chapter 2 the concept of a sensitivity analysis of a trial design was introduced for data anticipated to take a Normal form where the trial's sensitivity was assessed with respect to the variance estimate used in the calculations. For Normal data an upper estimate of the variance, estimated from the chi-squared distribution, can be used to investigate the sensitivity of the designed study to a plausibly extreme value for the variance.

In Chapter 3 it was highlighted how for a binary response the variance estimate, both for a proportion and an odds-ratio, does not conform to a chi-squared distribution except for large sample sizes. The consequence of this is that the chi-squared asymptotic assumption for the estimate of the variance cannot be used to assess the sensitivity of the study design.

For binary data however, it is the response rate on the control arm, p_A, usually estimated from a previous clinical study that is used to estimate a population response, to which the study design is sensitive to. This control response rate in turn feeds into the estimate of variance used in the calculations. Hence, any imprecision in the estimation of the control response rate will impact on the study design.

To investigate the effect the imprecision of the estimate of the control response rate will have on the study design a range of plausible values could be obtained through construction of a 95% confidence interval (using methods described in Chapter 3). From the two tails of this confidence interval a re-estimation of the variance could be made. The power could then be assessed using these new variance estimates through use of the following equation for proportional differences – (4.1.4) rewritten in terms of power

$$1-\beta = \Phi\left(\sqrt{\frac{n(p_A-p_B)^2}{(p_A(1-p_A)+p_B(1-p_B))}} - Z_{1-\alpha/2}\right), \qquad (4.1.9)$$

and the follow for odds-ratios (4.1.3 rewritten in terms of power)

$$1-\beta = \Phi\left(\sqrt{n(\log OR)^2\left[1-\sum_{i=1}^{2}\bar{p}_i^3\right]/6} - Z_{1-\alpha/2}\right). \qquad (4.1.10)$$

These calculations would assess the sensitivity of the study design to plausible values for the control response rate.

4.2.1.7. Worked Example

Suppose now, from the worked example given earlier, that the response rate on control was estimated from a study with 50 patients. A Wilson (non-continuity corrected) 95%

4.2.1.6. Sensitivity Analysis about the Estimates of the Population Effects Used in the Sample Size Calculations

In Chapter 2 the concept of a sensitivity analysis of a trial design was introduced for data anticipated to take a Normal form where the trial's sensitivity was assessed with respect to the variance estimate used in the calculations. For Normal data an upper estimate of the variance, estimated from the chi-squared distribution, can be used to investigate the sensitivity of the designed study to a plausibly extreme value for the variance.

In Chapter 3 it was highlighted how for a binary response the variance estimate, both for a proportion and an odds-ratio, does not conform to a chi-squared distribution except for large sample sizes. The consequence of this is that the chi-squared asymptotic assumption for the estimate of the variance cannot be used to assess the sensitivity of the study design.

For binary data however, it is the response rate on the control arm, p_A, usually estimated from a previous clinical study that is used to estimate a population response, to which the study design is sensitive to. This control response rate in turn feeds into the estimate of variance used in the calculations. Hence, any imprecision in the estimation of the control response rate will impact on the study design.

To investigate the effect the imprecision of the estimate of the control response rate will have on the study design a range of plausible values could be obtained through construction of a 95% confidence interval (using methods described in Chapter 3). From the two tails of this confidence interval a re-estimation of the variance could be made. The power could then be assessed using these new variance estimates through use of the following equation for proportional differences – (4.1.4) rewritten in terms of power

$$1-\beta = \Phi\left(\sqrt{\frac{n(p_A-p_B)^2}{(p_A(1-p_A)+p_B(1-p_B))}} - Z_{1-\alpha/2}\right), \qquad (4.1.9)$$

and the follow for odds-ratios (4.1.3 rewritten in terms of power)

$$1-\beta = \Phi\left(\sqrt{n(\log OR)^2\left[1-\sum_{i=1}^{2}\bar{p}_i^3\right]/6} - Z_{1-\alpha/2}\right). \qquad (4.1.10)$$

These calculations would assess the sensitivity of the study design to plausible values for the control response rate.

4.2.1.7. Worked Example

Suppose now, from the worked example given earlier, that the response rate on control was estimated from a study with 50 patients. A Wilson (non-continuity corrected) 95%

confidence interval for this point estimate would give the true value as being between 36.4% and 63.4%. Table 4.3a gives the sensitivity of the study design using the odds-ratio to calculate the sample size. One can see from this result that there is only a small loss in power of 4% if a response rate of 36.4% was observed and a nominal loss if 63.4% was observed. For both these calculations the effect was assumed fixed at 1.5.

Table 4-3. Sensitivity analysis for superiority worked example

a. Odds-Ratio scale

		95% Confidence Interval	
	Observed	Lower	Upper
Control Response	0.50	0.364	0.634
Investigative Response	0.40	0.276	0.536
Power	90%	86%	89%

b. Proportional scale

		95% Confidence Interval	
	Observed	Lower	Upper
Control Response	0.50	0.364	0.634
Investigative Response	0.40	0.264	0.534
Power	90%	94%	90%

c. Proportional scale – same effects as the odds-ratio

		95% Confidence Interval	
	Observed	Lower	Upper
Control Response	0.50	0.364	0.634
Investigative Response	0.40	0.276	0.536
Power	90%	86%	89%

Table 4.3b gives the equivalent calculations assuming the effect was fixed at an absolute difference of 10%. One can see from these results that if a lower or higher response rate than expected was observed the power would actually increase. The reason for this, recalling Chapter 3, is that the maximum value the variance could possibly be on the proportional scale is when the response rate is expected to be in the middle of the range. This is a situation one anticipates to observe here (although the pooled two group variance is a little different as it accounts for the anticipated investigative response rate).

As was discussed in Chapter 3 the odds-ratio has the property that the same fixed odds-ratio will equate to different differences on the proportional scale – dependent on the anticipated control response. Table 4.3b assumed that no matter what the prevalence the same effect (10%) would be of interest whilst Table 4.2a, by default altered the effects of interest on the proportional scale (with a fixed OR). Table 4.3c repeats the calculations of 4.3b but using the same effects on the proportional scale as used in Table 4.3a. This table concurs with Table 4.3a. The issues of different responses of

interest according to the control prevalence will be discussed in greater detail in the sections on non-inferiority and equivalence trials.

The worked example here was a special case in that the response was anticipated to be toward the middle of the range. Suppose though that the control response was expected to be 20% and this was estimated from study with 50 patients on the control arm. The sample size required for an odds-ratio of 1.5 is 900 patients per arm. The equivalent calculation on the proportional scale gives a sample size of 917 (a control response rate of 20% and an odds-ratio of 1.5 equates to an absolute difference of 5.7%).

Table 4-4. Sensitivity analysis for superiority worked example

a. Odds-ratio scale

		95% Confidence Interval	
	Observed	Lower	Upper
Control Response	0.200	0.112	0.330
Investigative Response	0.143	0.078	0.247
Power	90%	71%	97%

b. Proportional scale

		95% Confidence Interval	
	Observed	Lower	Upper
Control Response	0.200	0.112	0.330
Investigative Response	0.143	0.055	0.273
Power	90%	99%	76%

c. Proportional scale – same effects as the odds-ratio

		95% Confidence Interval	
	Observed	Lower	Upper
Control Response	0.200	0.112	0.330
Investigative Response	0.143	0.078	0.247
Power	90%	70%	98%

Table 4.4 gives the same sensitivity analysis as conducted with Table 4.3. One can see here one gets quite markedly different answers to Table 4.3 with the proportional scale being quite sensitive to the assumptions around active response rate.

What these examples highlight is the complexity of investigating the sensitivity of study with an uncertain response rate. The sensitivity of the design varies according to the anticipated control response rate. The issues raised in this section will be re-addressed throughout this chapter.

4.2.1.8. Optimising the Estimates of the Population Effects

As described earlier in this chapter, to investigate the sensitivity of a study with binary data it is critical to estimate accurately the response rate on control, p_A, as this control response rate feeds into the variance estimate. If one has several clinical investigations from which one can obtain an estimate of the control response rate then an overall estimate of the response is required. To do this one could follow meta-analysis methodologies [Whitehead and Whitehead, 1991]. To obtain an overall estimate across several studies one could use

$$p_s = \frac{\sum_{i=1}^{k} w_i p_i}{\sum_{i=1}^{k} w_i}, \qquad (4.1.11)$$

where p_s is an estimate of the overall response, p_i is an estimate of the response from study i, w_i is the reciprocal of the variance from study i ($w_i = 1/\text{var}(p_i)$) and k is the number of studies. Hence, define,

$$p_i \sim N(p_s, w_i^{-1}), \qquad (4.1.12)$$

and thus

$$\sum_{i=1}^{k} w_i p_i \sim N\left(p_s \sum_{i=1}^{k} w_i, \sum_{i=1}^{k} w_i \right), \qquad (4.1.13)$$

and hence overall one can define

$$p_s = \frac{\sum_{i=1}^{k} w_i p_i}{\sum_{i=1}^{k} w_i} \sim N(\pi, \text{var}(\pi)). \qquad (4.1.14)$$

where π is the control response rate. The variance for p_s is defined as $p_s = 1/\sum_{i=1}^{k} w_i$ and consequently a 95% confidence interval for the overall estimate can be obtained from

$$p_s \pm Z_{1-\alpha/2} \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}}. \qquad (4.1.15).$$

Note that the methodology applied here is that of fixed effects meta-analysis. In the dissertation to date what has been accounted for is the variation that arises as a result of pure sampling error. Random trial to trial variability in the "true" control group rate has not been investigated. The approaches described in this section can allow one to undertake this investigation.

One could apply a random effects approach by replacing w_i with w_i^* where w_i^* comes from [Whitehead and Whitehead, 1991]

$$w_i^* = (w_i^{-1} + \tau^2)^{-1},$$

where τ is defined as

$$\tau^2 = \frac{\sum_{i=1}^{k} w_i (p_i - p_s)^2 - (k-1)}{\sum_{i=1}^{k} w_i - \left(\sum_{i=1}^{k} w_i^2 \Big/ \sum_{i=1}^{k} w_i \right)}.$$

Simply, τ^2 can crudely be thought of as

$$\tau^2 = \frac{\text{Variation in the treatment difference between groups}}{\text{Variation in the variation between groups}}.$$

If $\tau^2 = 0$ then the weighting for the fixed effect analysis is used.

The corresponding (random effects) confidence interval would be given by

$$p_s \pm Z_{1-\alpha/2} \sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}}.$$

The relative merits of fixed versus random effects meta analysis will not be discussed here. In this chapter the methodology applied will be that of fixed effects meta analysis.

One thing to highlight here however is that in context with this dissertation it is not so much random effects analysis but random effects planning that is of importance. Even if the decision is made that a fixed effects analysis will be undertaken - the issue of the control group remains. The fundamental assumption, when planning a trial, is that the true control group rates are the same from trial to trial and observed rates can only vary according due to sampling error. However, if this were the case, could one effectively say that one could use historical data to form an augmented control group? The very fact that concurrent controls are used is an admission of the fact that the belief is that true control group rates can vary from trial to trial. What this touches on, in fact, is the heterogeneity of trials, especially trials conducted sequentially over time or in different regions say. This will be discussed in greater detail in Chapter 6.

4.2.1.9. Worked Example

Table 4.5 gives the data from 8 different studies for the control response rate [Stampfer, Goldhaber, Yusuf et al, 1982]. As one can see the response rates vary between 8% and 27% across the different studies. The final two columns give the workings for calculations for w_i and $w_i p_i$ and hence the calculations for the overall estimates. From

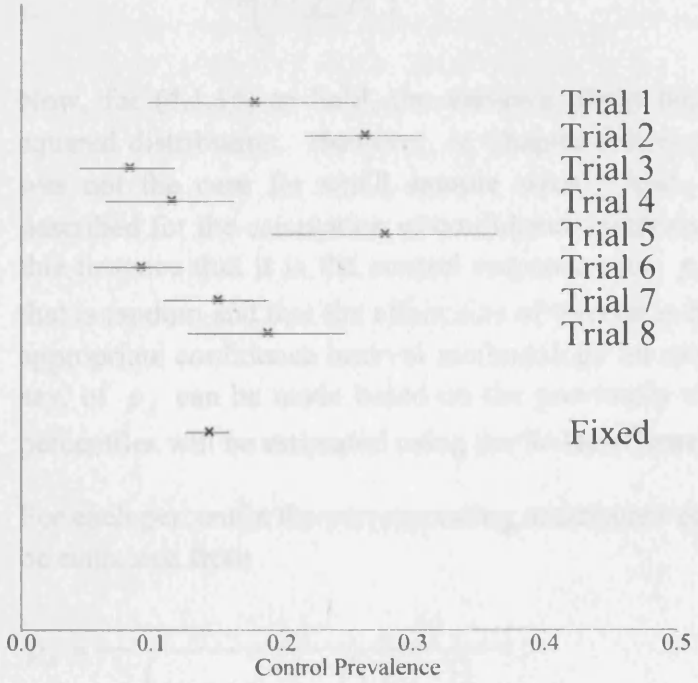
this one can estimate the overall response to be 14.3% with standard error 0.0086. Hence, the 95% confidence interval around the overall estimate is (0.126 to 0.160).

Table 4-5. Table of control data by individual study

	Control				
Trial	d	Total	p_i	w_i	$p_i w_i$
1	15	84	0.179	572.66	102.26
2	94	357	0.263	1840.44	484.60
3	17	207	0.082	2746.05	225.52
4	18	157	0.115	1546.72	177.33
5	29	104	0.279	517.18	144.21
6	23	253	0.091	3061.30	278.30
7	44	293	0.150	2295.89	344.78
8	30	159	0.189	1038.68	195.98
Total	270	1614		13618.91	1952.97

The response from each study and the overall response estimate are given in figure 4.1. There may be some evidence of heterogeneity across the studies used in this example. This may be because certain trials were sampled from "different" populations. As mentioned earlier the issue of heterogeneity across studies will be investigated in Chapter 6.

Figure 4-1. Plot of point estimates and confidence intervals for individual studies and overall



4.2.1.10. Calculations Taking Accounting of the Imprecision of the Estimates of the Population Effects Used in the Sample Size Calculations

4.2.1.11. Odds-Ratio

Earlier in this chapter sample size calculations were given for the case where the population effects were assumed known. However, when calculating a sample size the population effects are not usually known but are estimated from previous research – a study say of similar design where the control regimen was given.

In Chapter 2 it was shown that for Normally distributed data, where a sample estimate is used instead of the population variance, the expected power could be determined from a non-central t-distribution. Now if arguments for Normally distributed data could be generalised to data which take a binary form the sample size per group would be derived from

$$n = \frac{6\left[tinv(1-\beta, df, Z_{1-\alpha/2})\right]^2 / (\log OR)^2}{\left[1 - \sum_{i=1}^{2} \bar{p}_i^3\right]}, \qquad (4.1.16)$$

where *TINV(•,m,a)* denotes the (monotonically increasing) inverse function of the cumulative distribution of a non-central t distribution with m degrees of freedom and non-centrality parameter a. The degrees of freedom (d.f.) in the formula refers to the degrees of freedom about the variance estimate used in the sample size calculation.

For the case of the log(OR) remember the approximate result

$$Var\left(\log(OR)\right) = \frac{6}{n\left(1 - \sum_{i=1}^{2} \bar{p}_i^3\right)}. \qquad (4.1.17)$$

Now, for (4.1.16) to hold, the variance about the estimates should following a chi-squared distribution. However, in Chapter 3 it was empirically demonstrated that this was not the case for small sample sizes. Also, in Chapter 3 methodologies were described for the calculation of confidence intervals for a single proportion. Assume in this instance that it is the control response rate, p_A, estimated from a previous study that is random and that the effect size of interest is the odds-ratio and is fixed. Using an appropriate confidence interval methodology an estimate of the 1st, 2nd 3rd percentile, say, of p_A can be made based on the previously observed p_A. In this instance these percentiles will be estimated using the Wilson Score (not continuity corrected) method.

For each percentile the corresponding anticipated response on the investigative arm can be estimated from

$$p_B = \frac{1}{\exp\left(\log(OR) - \log\left(\frac{p_A}{1-p_A}\right)\right) + 1}, \qquad (4.1.18)$$

and an estimate of the variance from (4.1.17). If one took the average across all the percentiles then for a given sample size, n, and imprecision around the estimate of p_A the power can be estimated from

$$1-\beta = \frac{1}{0.998}\sum_{perc=0.001}^{0.998} 0.5\left[\begin{array}{l}\Phi\left(\sqrt{n(\log OR)^2\left[1-\sum_{i=1}^{2}\bar{p}_{perc}^{3}\right]/6}-Z_{1-\alpha/2}\right)\\ +\Phi\left(\sqrt{n(\log OR)^2\left[1-\sum_{i=1}^{2}\bar{p}_{(perc+0.001)}^{3}\right]/6}-Z_{1-\alpha/2}\right)\end{array}\right]. \qquad (4.1.19)$$

(4.1.19) can be iterated on n until the appropriate power has been reached.

4.2.1.12. Comparison of the Two Methods

A comparison of the two methods is given in Table 4.6. The sample sizes are estimated from (4.1.19) and (4.1.16). From empirical observation of table it seems that if the anticipated control response rate is 0.20 then (4.1.16) and (4.1.19) agree however, overall in comparison to (4.1.19) the results from (4.1.16) can be summarised as follows

0.10 - Under estimation

0.20 - Close agreement

0.30 - - Over estimation

Note there is a symmetry to the results such that the result for an anticipated response of 0.90 (not given) is similar to 0.10; 0.80 is similar to 0.20 and 0.50 to 0.70 are similar to 0.30 to 0.40.

The results from Table 4.6 are not surprising when one remembers what was observed in Chapter 3. First of all for small sample sizes the variance of the estimated log-odds-ratio is not well approximated by the chi-squared distribution. Second through the mid range of anticipated observed response rates (between 0.3 and 0.7) the variance is relatively flat and does not deviate much. Hence, the over estimation of the sample size of (4.1.16), compared to (4.1.19). Finally as the anticipated responses tend towards the limits (0,1) there are quite marked deviations in the variance depending the anticipated responses. Hence the under estimation of (4.1.16) compared to (4.1.19).

Table 4-6. Table of sample sizes for a fixed odds-ratio of 2, for different assumed control responses and degrees of freedom around the sample variance. Calculated using numerical methods and the non-central t-distribution

Control Response	Degrees of Freedom	Numerical Sample Size	Non-central t Sample Size
0.10	10	1694	807
	15	1061	738
	20	902	707
	25	829	688
	50	713	653
	75	680	642
	100	665	637
0.20	10	455	433
	15	407	396
	20	387	379
	25	375	369
	50	354	351
	75	347	345
	100	344	342
0.30	10	286	313
	15	271	287
	20	263	274
	25	259	267
	50	250	254
	75	247	249
	100	246	247
	250	243	244
	500	242	242
0.40	10	224	259
	15	216	237
	20	212	227
	25	210	221
	50	205	210
	75	203	206
	100	202	205

4.2.1.13. Proportional Difference

The equivalent formula to (4.1.16) for calculations done on the proportional scale is

$$n = \frac{[tinv(1-\beta, df, Z_{1-\alpha/2})]^2 (p_A(1-p_A) + p_B(1-p_B))}{(p_A - p_B)^2}. \qquad (4.1.20)$$

Similar to the case with the odds-ratio, for (4.1.20) to hold the variance about the estimates should follow a chi-squared distribution. However, as with the odds-ratio, it was shown in Chapter 3 that empirically this assumption does not hold for small sample sizes.

Remember that, for the difference in proportions, $p_A - p_B$, the variance is defined as

$$Var(p_A - p_B) = \frac{p_A(1-p_A)}{n} + \frac{p_B(1-p_B)}{n}. \qquad (4.1.21)$$

As described for odds-ratio using the Wilson Score method percentiles for p_A can be estimated from previously observed p_A. Now assuming the effect size, $p_A - p_B$, is fixed then for each percentile the corresponding anticipated response on the investigative arm can be estimated from

$$p_B = p_A + Effect \qquad (4.1.22)$$

and an estimate of the variance made from (4.1.21). Correspondingly an estimate of the power for a given n and imprecision about p_A can be made from

$$1-\beta = \frac{1}{0.998}\sum_{p_{(pres)}=0.001}^{0.998} 0.5\left[\Phi\left(\sqrt{\frac{n(p_A-p_B)^2}{\left(p_{(pres)_A}(1-p_{(pres)_A})+p_{(pres)_B}(1-p_{(pres)_B})\right)}} - Z_{1-\alpha/2}\right) + \Phi\left(\sqrt{\frac{n(p_A-p_B)^2}{\left(p_{(pres-0.001)_A}(1-p_{(pres-0.001)_A})+p_{(pres-0.001)_B}(1-p_{(pres-0.001)_B})\right)}} - Z_{1-\alpha/2}\right)\right]. \qquad (4.1.23)$$

The sample size can be estimated through iteration.

4.2.1.14. Comparison of the Two Methods

A comparison of the two methods is given in Table 4.7. The final two columns give the sample size calculated using (4.1.20) and from (4.1.23) respectively.

From empirical observation of the table it seems that if the anticipated control response rate is 0.05 then (4.1.20) and (4.1.23) agree. For the other control responses (4.1.20) over estimates the sample size. Responses over 0.25 are not given as between 0.30 and 0.50 there is no advantage in accounting for the imprecision according to (4.1.23) which returns the same sample size for all response. Table 4.7 can be summarised as follows

0.05	Close agreement
0.10 or 0.25	Over estimation
0.30 to 0.50	Over estimation with (4.1.23) returning the same sample size as standard formula

Again the results from Table 4.7 are not surprising when one remembers what was observed in Chapter 3. First of all as with the log-odds-ratio for small sample sizes the variance around a proportion does not follow the chi-squared asymptotic form. Second through the mid range of anticipated observed response rates (between 0.3 and 0.7) the variance is relatively flat and does not deviate much. Also the peak possible variance is for when the anticipated response it around 0.50. Hence accounting for the imprecision around the estimates of the population effects has no affect in the mid range. Finally as the anticipated responses tend towards the limits (0,1) there are quite marked deviations

in the variance depending the anticipated responses hence the variation in the sample size when accounting for the imprecision around the sample variance.

Table 4-7. Table of sample sizes for a fixed proportional difference of 0.10, for different assumed control responses and degrees of freedom around the sample variance. calculated using numerical methods and the non-central t-distribution

Control Response	Degrees of Freedom	Numerical Sample Size	Non-Central t Sample Size
0.05	10	234	239.1
	20	215	209.4
	30	207	200.5
	40	202	196.2
	50	199	193.6
	100	192	188.7
0.10	10	298	341.6
	20	285	299.2
	30	279	286.4
	40	276	280.2
	50	274	276.6
	100	269	269.6
	250	266	265.4
	500	264	264.0
0.15	10	350	430.5
	20	344	377.0
	30	340	360.9
	40	339	353.1
	50	337	348.6
	100	335	339.6
0.20	10	393	505.6
	20	392	442.8
	30	392	423.9
	40	391	414.8
	50	391	409.4
	100	390	398.9

4.2.1.15. Worked Example

An investigator wishes to design a study where the response anticipated on the control therapy was 20%. The effect of interest is an odds-ratio of 2.0 in favour of the control therapy (i.e. the aim is to reduce the number of events) and the investigator wishes to design the study with Type I and II errors fixed at 5% and 10% respectively. From Table 4.1 one can see that the sample size required would be 333 patients per arm of the trial.

Now suppose this estimate of the control response rate came from a trial with 50 patients receiving control. To allow for the imprecision in the estimate of the control response rate the sample size (from Table 4.6) would need to increase to 354 patients

With a response rate of 20% anticipated on control, an odds-ratio of 2.0 would equate to reducing the investigative response rate to 11.11% or an 8.89% reduction. From (4.1.4) one can see that the sample size required to detect this difference is 344 patients. When one allows for the imprecision in the control response rate estimate the sample size (from (4.1.23)) is increased to 355.

4.2.1.16. Calculations Taking Accounting of the Imprecision of the Estimates Used in the Sample Size Calculations – Bayesian Methods

In Chapter 2 there was a discussion on the use of Bayesian methods for sample sizes for non-inferiority and equivalence studies where there was uncertainty about the mean difference. The calculations were for data anticipated to take a Normal form. For this situation there is no great practical application for this work however.

If the primary endpoint is a binary response then it is the uncertainty in the mean (proportional) response which adversely affects sample size calculations. The context now is to interrogate sample sizes where a superiority study is be planned and a control response, p_A, had previously been observed. In the prospective trial being designed inference is to be made about the 'true' difference $\pi_A - \pi_B$. In context with the problem here the effect size (whether it be an odds-ratio or a proportional difference), is assumed known so that the variance for the odds-ratio and proportional difference can be estimated from (4.1.17) and (4.1.21) respectively as before.

For the given sample size what needs to be determined is the probability of observing a given control response, p_{A_0}, or greater for θ given that p_{A1} has already been observed i.e. $\Pr ob(\theta > p_{A_0} | p_{A1})$.

Note, similar to Normal data in Chapter 2, the methodology described in the subsections, (4.1.1.11) and (4.1.1.14), could be considered to be sample calculations calculated under a Bayesian framework but with a non-informative prior distribution for θ.

For inference about an unknown binary parameter, θ, what one is interested in is how would our belief about θ change. If the prior is expressed in the density $p(\theta)$ and if subsequently data x are observed then the posterior distribution is expressed in the density, $p(\theta|x)$, where the Bayes rule for densities is

$$p(\theta|x) \propto \lambda(x|\theta)\, p(\theta), \qquad (4.1.24)$$

where $\lambda(x|\theta)$ is the likelihood function.

For binary data the Beta distribution can be used for the prior responses such that

$$PROBBETA(p_A, a, b) \propto p_A^{a-1}(1-p_A)^{b-1}, \quad (4.1.25)$$

where $PROBBETA(\bullet)$ is defined as a cumulative density distribution for a beta distribution. The Bayesian updating rules are now described. Although not directly comparable this chapter draws on the work of Johnson, Su, Gardner et al [2004].

4.2.1.17. Prior Response

Prior values for $PROBBETA(p_{percentile}, a_0, b_0)$ (and the corresponding $P_{perc} = BETAINV(perc, a_0, b_0)$) could be derived as follows. For an informative prior one could use the mode (or most likely value) and a percentile to build a prior. For a Beta distribution the mode is defined by

$$m = \frac{a_0 - 1}{a_0 + b_0 - 2}.$$

Hence, the a_0 (and consequently b_0) could be derived from

$$p_{percentile} = BETAINV(percentile, a_0, [a_0 - 1]/m - a_0 + 2), \quad (4.1.26)$$

if a percentile for the control response can be postulated.

If one wished to use a non-informative prior then a Jeffrey's prior could be used such that

$$P_{perc} = BETAINV(perc, 0.5, 0.5). \quad (4.1.27)$$

This Jeffrey's prior has the advantage of being invariant with respect to transformations.

4.2.1.18. Anticipated Response

The anticipated control response (and consequent variance) is defined as p_{A_1}. This value is taken from an objective value observed in a previous clinical trial. The control response equates to an observed number of successes (a_1) and failures (b_1).

4.2.1.19. Posterior Response

With the anticipated and prior responses the posterior distribution can be calculated from the following result

$$P_{perc} = BETAINV(perc, a_1 + a_0, b_1 + b_0). \quad (4.1.28)$$

These values for $P_{new} = BETAINV(perc, a_1 + a_0, b_1 + b_0)$ can be used in (4.1.19) and (4.1.23) to obtain estimates of the sample size for an odds-ratio and a proportional difference respectively accounting for the imprecision in the variance estimate.

4.2.1.20. Worked Example

Repeating the earlier example where an investigator wished to design a study where the response anticipated on the control therapy was 20%. The effect size of interest is an odds-ratio of 2.0 in favour of the control therapy with the investigator wishing to design the study with Type I and II errors fixed at 5% and 10% respectively.

Now, again, suppose this estimate of the control response rate came from a trial with 50 patients receiving control and the investigator wished to allow for this imprecision in the estimate of the control response rate in the estimation of the sample size.

If initially a non-informative prior was used then using (4.1.27) in (4.1.19) the sample size is estimated to be 354 patients. The sample size is as calculated previously when allowing for the imprecision in the sample variance.

Imagine the investigator was sceptical as to the control response being as high as 20% such that the belief was that the most likely response was 15% with at least 90% certainty that it was greater than 10%. From (4.1.26) estimates of a_0 and b_0 of 7.899 and 40.094 are obtained. Hence, now using (4.1.28) in (4.1.19) the estimate of sample size is now 379.

Now suppose the investigator is more optimistic about the control response believing the most like response to be 25% and is at least 90% certain that it was greater than 20%. From (4.1.26) estimates of a_0 and b_0 of 25.048 and 73.144 are obtained. Hence, now using (4.1.28) in (4.1.23) a sample size estimate of 297 is calculated. This is less than the sample size calculated not allowing for imprecision in the variance.

Similar calculations could be done if calculations were based around an absolute difference in the responses. Here though (in terms of the variance) an optimistic prior would be one where the response is lower than 20% and pessimistic prior would be where the response is higher than 20%.

4.2.2. Cross-over Trials

When the data are paired, such as in a cross-over trial, there are two main summary measures, the difference in proportions and the odds-ratio, that may be used. This section will concentrate on these two summary measures in considering sample size calculations. Also, the sample size calculations will depend on whether the effect of period will or will not be allowed for in the final analysis. The methodologies will now be discussed in detail.

4.2.2.1. Ignoring Period

4.2.2.2. Analysis

For a cross-over trial with binary data the data could be summarised as per Table 4.8 and analysed by the McNemar test

$$\frac{(n_{10}-n_{01})^2}{n_{10}+n_{01}} \sim \chi_1^2 \quad ,$$

where n_{10} and n_{01} are the number of responses expected in cells '10' and '01'. The data in the final column and final row give the overall responses for each treatment. These overall responses are the outcomes one may consider to be expected in a parallel group study.

Table 4-8. Summary table of hypothetical cross-over trial

		Treatment B		
		1	0	
Treatment A	1	n_{11}	n_{10}	n_{A1}
	0	n_{01}	n_{00}	n_{A0}
		n_{B1}	n_{B0}	n

In a cross-over trial only discordant responses are of interest for statistical comparisons i.e. those subjects who respond '10' or '01'. A large proportion of the data are thus discarded in constructing a statistical test, as the test is conditional on subjects being discordant. This is quite intuitive though, as in a superiority trial concordant responses concur with the null hypothesis of no treatment differences. Thus what one is determining is whether for those subjects who only respond to one treatment, this response is more likely to be in favour of one treatment over the other.

4.2.2.3. Sample Size Estimation

4.2.2.4. Population Effects Assumed Known

Table 4.8 can be rewritten in terms of proportions as per Table 4.9, where $\lambda_{10} = n_{10}/n$, $\lambda_{01} = n_{01}/n$, $\lambda_{11} = n_{11}/n$ and $\lambda_{00} = n_{00}/n$ and $P_A = n_{A1}/n$ and $P_B = n_{B1}/n$ and the trial can be summarised with an odds ratio defined as

$$\psi = \frac{\lambda_{10}}{\lambda_{01}}. \qquad (4.1.29)$$

Table 4-9. Summary table of hypothetical cross-over trial

		Treatment B		
		1	0	
Treatment A	1	λ_{11}	λ_{10}	P_A
	0	λ_{01}	λ_{00}	$1-P_A$
		P_B	$1-P_B$	1

This odds-ratio is a conditional summary statistic, using just the discordant responses. A conditional odds-ratio can be difficult to interpret. To assist in the interpretation the odds-ratio can be approximated from the marginal totals [Royston, 1993]

$$\text{Odds-ratio}=\psi \approx \frac{p_A(1-p_B)}{p_B(1-p_A)}. \qquad (4.1.30)$$

where $\lambda_{10} \approx p_A(1-p_B)$ and $\lambda_{01} \approx p_B(1-p_A)$. Thus, the conditional odds-ratio for a cross-over trial can be interpreted in terms of the odds-ratio from a parallel group study (approximated from the marginal proportions). This is of particular use in the calculation of sample sizes, as marginal totals could be used to estimate the conditional odds-ratio, which in turn can be used to estimate the discordant sample size.

The discordant sample size, n_d, for a cross-over trial can be derived from [Royston, 1993; Julious, Campbell and Altman, 1995; Connett, Smith and McHugh, 1987; Fleiss and Levin, 1981; Schesselman, 1982]

$$n_d = \frac{\left(Z_{1-\alpha/2}(\psi+1)+2Z_{1-\beta}\sqrt{\psi}\right)^2}{(\psi-1)^2}, \qquad (4.1.31)$$

which has shown to perform well in simulations [Julious and Campbell, 1998]. To calculate the total sample size, the discordant sample size is divided by the proportion expected to be discordant [Julious, Campbell and Altman, 1995; Connett, Smith and McHugh, 1987] i.e.

$$N_c = \frac{n_d}{\lambda_{01}+\lambda_{10}}. \qquad (4.1.32)$$

There are alternative formulae, which do not require a two-stage approach to calculate the total sample size [Julious, Campbell and Altman, 1995; Connett, Smith and McHugh, 1987; Conner, 1987; Miettinen, 1968]. However, this chapter though will concentrate on the two-stage approach.

The equivalent sample size for one arm in a parallel group trial, N_{pg}, can be estimated from (4.1.33) now defined slightly differently as [Campbell, Julious and Altman, 1995; Whitehead, 1993]

$$N_{pg} = \frac{6\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 / (\log OR)^2}{\left[1 - \sum_{i=0}^{1} \bar{p}_i^3\right]}, \qquad (4.1.33)$$

where similarly to (4.1.3) $\bar{p}_0$ and $\bar{p}_1$ are the responses across treatments for outcomes 0 and 1 such that $\bar{p}_1 = (p_A + p_B)/2$ and $\bar{p}_1 = 1 - \bar{p}_0$.

Similar to the situation described earlier in the chapter for the sample size formulae for odds-ratios and proportional differences, on the face of it (4.1.31) is quite dissimilar to (4.1.33). However, (4.1.31) can be re-written as

$$N_c = \frac{\left(Z_{1-\alpha/2}(\lambda_{10} + \lambda_{01}) + 2Z_{1-\beta}\sqrt{\lambda_{10}\lambda_{01}}\right)^2}{(\lambda_{10} + \lambda_{01})(\lambda_{10} - \lambda_{01})^2}, \qquad (4.1.29)$$

and in turn re-writing λ_{10} and λ_{01} in terms of the marginal totals ($\lambda_{10} \approx p_A(1-p_B)$ and $\lambda_{01} \approx p_B(1-p_A)$) (4.1.34) can be approximated by

$$N_c \approx \frac{\left(Z_{1-\alpha/2}(p_A(1-p_B) + p_B(1-p_A)) + 2Z_{1-\beta}\sqrt{p_A(1-p_B)p_A(1-p_B)}\right)^2}{(p_A(1-p_B) + p_B(1-p_A))(p_A(1-p_B) - p_B(1-p_A))^2}. \qquad (4.1.35)$$

Also, through the following results $p_A(1-p_B)p_A(1-p_B) \approx \bar{p}_0^2(1-\bar{p}_0)^2$ and $p_A(1-p_B) + p_B(1-p_A) \approx 2\bar{p}(1-\bar{p})$ (4.1.30) can be re-written as

$$N_c \approx \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 2\bar{p}_0^2(1-\bar{p}_0^2)^2}{\bar{p}_0(1-\bar{p}_0)(p_A - p_B)^2}. \qquad (4.1.36)$$

Returning back to the case of a parallel group trial where the odds-ratio (OR) is defined as $OR = (p_A(1-p_B))/(p_B(1-p_A))$ and remember

$$\log(OR) \approx \frac{p_A - p_B}{\bar{p}_0(1-\bar{p}_0)}, \qquad (4.1.37)$$

$$\frac{6}{\left(1 - \sum_{i=0}^{1} \bar{p}_i^3\right)} = \frac{2}{\bar{p}_0(1-\bar{p}_0)}. \qquad (4.1.38)$$

Substituting (4.1.37) and (4.1.38) back into (4.1.36) one gets

$$N_c \approx \frac{6\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 / (\log OR)^2}{\left[1 - \sum_{i=0}^{1} \bar{p}_i^3\right]} = N_{pg},$$

and the sample size formula for one arm in a parallel group study. Thus, the sample size required for a cross-over trial is approximately equivalent to that for one arm of a parallel group trial. An alternative way of phrasing this would be to say that the sample

size required is half that required in total for a parallel group trial. Table 4.10 also gives empirical evidence of this fact, giving the sample size required for various expected outcome responses for a given treatment and odds-ratios using (4.1.33) and (4.1.34). These results demonstrate that if the parallel group formula, (4.1.34), was used, slightly smaller estimates of the sample size for a cross-over trial would be obtained compared to (4.1.33) but only marginally so. Practically they give the same sample size.

Table 4-10. Sample size estimates for a cross-over trial (n_c) and one arm of a parallel group trial (n_{pg}) for various expected outcome responses for a given treatment (p_A) and odds-ratios for a two sided type I error rate of 5% and 90% power

	Odds-Ratio											
	1.25		1.50		1.75		2.00		3.00		4.00	
P_A	n_c	n_{pg}	n_c	n_{pg}	n_c	n_{pg}	n_c	n_{pg}	n_c	n_{pg}	n_c	n_{pg}
0.20	2824	2819	905	900	499	494	339	333	154	146	106	98
0.40	1804	1803	561	560	302	300	201	200	87	87	58	56
0.60	1726	1725	518	517	270	270	176	176	71	70	46	45
0.80	2471	2468	711	709	359	356	226	224	83	71	50	49

Intuitively these results seem reasonable as the analysis one is undertaking in a cross-over trial, the McNemar test, is simply a one sample χ^2-test. For the analogous paired t-test, as highlighted in Chapter 2, the estimated sample size from one arm of a parallel group study provides an approximately equivalent sample for the paired sample size. It is good therefore to see that the same rationale can be applied to binary data.

The practical application of this result is that when designing a clinical trial one could use the marginal effects expected for the respective treatments and consequently the effect sizes anticipated if the trial was a parallel group investigation. These effects could be then used in the parallel group formula – taking the one arm sample size to be the total sample size. Working with the marginal totals may make it easier to formulate effects and consequently trials should be easier to design.

From now on in this chapter the approach of using the sample size formula for one arm of a parallel group trial as the total sample for a cross-over trials will be the approach applied and no great detail will go into the discussion of sample sizes for cross-over trials.

Note that although the arguments here concentrated on odds-ratios, as was demonstrated earlier in the chapter one can move interchangeably between odd-ratios and proportional differences and hence the arguments can be generalised to proportional differences with respect to cross-over trials.

Note also that the conditional odds-ratio is not the same as the marginal odds ratio but can be an approximation to "all things being equal". This issue is discussed again later in the chapter.

4.2.2.5. Worked Example

An investigator wishes to design a study where the marginal response anticipated on the control therapy is 40%. The effect if interest is 2.0 in favour of the control therapy and the investigator wishes to design the study with Type I and II errors fixed at 5% and 10% respectively.

An anticipated control response of 40% and an odds-ratio of 2.0 would equate to a response of 25% on the investigative therapy. Hence, the marginal responses, as per Table 4.11, can be completed, as well as the remaining entries in the table through multiplying the marginal totals. From this table it is evident that the odds-ratio defined through (4.1.29) and (4.1.30) is now the same.

Table 4-11. Summary table of anticipated responses for worked example

		Investigative		
		1	0	
Control	1	0.10	0.30	0.40
	0	0.15	0.45	0.60
		0.25	0.75	1

From Table 4.10 one can see that if (4.1.31) the sample size required would be 201 in total and if (4.1.33) were used the sample size would be 200. Practically this is the same.

If one wished to base the sample size purely on the discordant sample size, recruiting until the discordant sample size is reached, then the sample size would be (using the 200 per arm result) 200x(0.30+0.15) or 90 patients.

4.2.2.6. Alternative Sample Size Formulae

One aspect of designing cross-over trials with binary endpoints is the diverse array of alternative formulae that can be applied. Upon inspection, however, this diversity is down not to any technical difference in results but to definition of effect to be inserted into the sample size formula.

Here, as an aside, this chapter will briefly describe the different alternative formulae for the discordant sample size (two stage approach) and total sample size.

4.2.2.7. Discordant Sample Size

To calculate the discordant sample size Schesselman [1982] gives the following formula

$$n_d = \frac{\left(0.5Z_{1-\alpha/2} + Z_{1-\beta}\sqrt{P(1-P)}\right)^2}{(P-0.5)^2}, \tag{4.1.39}$$

where $P = \frac{\psi}{1+\varphi}$, (4.1.40)

and φ is the conditional odds-ratio for binary data defined as

$$\psi = \frac{\lambda_{10}}{\lambda_{01}}. \tag{4.1.41}$$

Connett, Smith and McHugh [1987] give a different format to Schesselman's equation simply by putting (4.1.40) into (4.1.39) to get [Royston, 1993; Fleiss and Levin, 1988]

$$n_d = \frac{\left(0.5Z_{1-\alpha/2} + Z_{1-\beta}\sqrt{\left(\frac{\varphi}{\psi+1}\right)\left(\frac{\varphi+1}{\psi+1} - \frac{\psi}{\psi+1}\right)}\right)^2}{\left(\frac{2\psi}{2(\psi+1)} - \frac{\psi+1}{2(\psi+1)}\right)^2},$$

$$n_d = \frac{\left(Z_{1-\alpha/2}(\psi+1) + 2Z_{1-\beta}\sqrt{\psi}\right)^2}{(\psi-1)^2}, \tag{4.1.42}$$

and (4.1.31) given earlier in this chapter. The results above are comparable to the formula given by Julious and Campbell [1998]

$$n_d = \frac{4\varphi\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{(\psi-1)^2} + \frac{Z_{1-\alpha/2}^2}{2}. \tag{4.1.43}$$

4.2.2.8. Total Sample Size

As discussed earlier in the chapter, to calculate the total sample size one can just divide the discordant sample size by the expected discordant proportion to obtain a sample size, termed a conditional approach i.e.

$$n_t = \frac{n_d}{\lambda_{01} + \lambda_{10}}. \tag{4.1.44}$$

Alternatively one can adopt a number of unconditional approaches such as that of Miettenen [1968]

$$n = \frac{\left(Z_{1-\alpha/2}\sqrt{r} + Z_{1-\beta}\sqrt{\left(r - \delta^2(3+r)\right)}\right)^2}{\delta^2}, \tag{4.1.45}$$

where $r = \lambda_{10} + \lambda_{01}$ and $\delta = \lambda_{10} - \lambda_{01}$. The Miettinen formula is not dissimilar to the formula given by Conner [1987]

$$n_t = \frac{\left(Z_{1-\alpha/2}\sqrt{r} + Z_{1-\beta}\sqrt{r-\delta^2}\right)^2}{\delta^2}, \qquad (4.1.46)$$

where r and δ are as given for Miettinen. Connett, Smith and McHugh [1987] gives an alternative format to that for Conner as $r = \lambda_{10} + \lambda_{01} = (\psi + 1)\lambda_{01}$ and $\delta = \lambda_{10} - \lambda_{01} = (\psi - 1)\lambda_{01}$, which if one substitutes into (4.1.41) gives

$$n_t = \frac{\left(Z_{1-\alpha/2}\sqrt{(\psi+1)} + Z_{1-\beta}\sqrt{(\psi+1) - (\psi-1)^2\lambda_{01}}\right)^2}{(\psi-1)^2\lambda_{01}}. \qquad (4.1.47)$$

Finally, multiply (4.1.47) top and bottom by $(\psi + 1)$ and one obtains the formula of Julious, Campbell and Altman [1999]

$$n_{tot} = \frac{[Z_{1-\alpha/2}(\psi+1) + Z_{1-\beta}\sqrt{(\psi+1)^2 - (\psi-1)^2 r}]^2}{(\psi-1)^2 r}. \qquad (4.1.48)$$

4.2.2.9. Sensitivity Analysis about the Estimates of the Population Effects Used in the Sample Size Calculations

Following on from the arguments on equating the one arm sample size of a parallel trial with that of the total sample size of a cross-over study, the methodology described earlier in this chapter for parallel group studies can be adapted to assess the sensitivity of a cross-over trial.

To investigate sensitivity of the study design to the imprecision of the estimate of control marginal response rate a range of plausible values could be constructed through a 95% confidence interval. The power could then be used assessed through for the two tails of this confidence interval, with the effect size fixed, by using the following - (4.1.28) rewritten in terms of power

$$1 - \beta = \Phi\left(\sqrt{n(\log OR)^2\left[1 - \sum_{i=0}^{1}\bar{p}_i^3\right]/6} - Z_{1-\alpha/2}\right). \qquad (4.1.49)$$

4.2.2.10. Worked Example

Suppose the marginal response rate on control of 40% given in the earlier worked example was estimated from a study with 40 patients. A Wilson (non continuity corrected) confidence interval for this point estimate would give the true value as being between 26.3% and 55.4% which would equate to powers of 80.2% and 93.3% respectively for the fixed sample size of 200 patients and odds-ratio of 2.

4.2.2.11. Calculations Taking Accounting of the Imprecision of the Estimates of the Population Effects Used in the Sample Size Calculations

As with assessing the sensitivity of a study, to calculate the total sample size of a cross-over study to account for the imprecision in the variance estimate used in the sample size calculations the results from the parallel group case can be extended to give the following result

$$1-\beta = \frac{1}{0.998}\sum_{perc=0.001}^{0.998} 0.5\left[\begin{array}{l}\Phi\left(\sqrt{n(\log OR)^2\left[1-\sum_{i=0}^{1} p_{perc}^3\right]}/6 - Z_{1-\alpha/2}\right) \\ +\Phi\left(\sqrt{n(\log OR)^2\left[1-\sum_{i=0}^{1} p_{(perc-0.001)}^3\right]}/6 - Z_{1-\alpha/2}\right)\end{array}\right], \qquad (4.1.50)$$

which can be iterated on n until the appropriate power has been reached.

4.2.2.12. Worked Example

Repeating the example gain given earlier of an anticipated marginal control response of 40% and odds-ratio of 2. If this response rate was estimated from 25 patients from a previous study then from Table 4.6 the total sample size is 212 (7% more than before).

4.2.2.13. Calculations Taking Accounting of the Imprecision of the Estimates Used in the Sample Size Calculations – Bayesian Methods

Bayesian methods described for parallel group trials can also be extended to studies with a cross-over design. A posterior distribution for a control response can be estimated and (4.1.50) can be used for sample size estimation.

4.2.2.14. Accounting for Possible Period Effects

4.2.2.15. Analysis

As well as ignoring concordant data, the McNemar test ignores the fact that subjects were assigned to different sequences, i.e. either AB or BA for a two period cross-over trial, and thus ignores any possible period effect which may exist. To allow for any possible period effect Table 4.8 can be re-written as sequence differences as in Table 4.12. The numbers in Table 4.12 can, in turn, be re-written in terms of Table 4.8 as $a_1+a_2=n_{10}$, $b_1+b_2=n_{01}$ and $n_{AB}+n_{BA}=n_d$.

This approach is analogous to the period adjusted t-test [Senn, 1993]. Sequence differences can be used to give a period adjusted estimate of the odds ratio, by taking the log-odds-ratio for sequence B-A away from A-B and dividing by 2

$\text{Log}\psi = (\log\psi_{AB} - \log\psi_{BA})/2 = (\log a_1/b_1 - \log b_2/a_2)/2 = 0.5\log(a_1a_2/b_1b_2)$

$$= 0.5\log OR_p. \qquad (4.1.51)$$

Where OR_p is the period adjusted odds-ratio. From (4.1.51) it is therefore evident that the non-period odds-ratio is equivalent to the square rooted odds-ratio from the period adjusted analysis. Thus, $\psi = \sqrt{OR_p}$ and hence a test statistic for the period adjusted test can be derived

$$\frac{(\log\psi)^2}{\text{var}\log\psi} \sim \chi_1^2 , \qquad (4.1.52)$$

where (4.1.52) is asymptotically equivalent to the McNemar test as well as to alternative period adjusted tests such as the Mainland-Gartt test and the Prescott test [Senn, 1993].

Table 4-12. Summary table of period adjusted analysis of hypothetical cross-over trial

	Treatment Difference		
Sequence Difference	-1	1	Total
A-B	a_1	b_1	n_{AB}
B-A	b_2	a_2	n_{BA}

The period adjusted approach described in this paper is an extension of the two group described by Whitehead [1993] and by McCullagh [1980]. The advantage of this approach is that it gives a measure of treatment effect, the odds-ratio, which is easily interpretable. Senn describes a similar approach, although that approach includes concordant data in the analysis [Senn, 1993]. The period-adjusted analysis can be undertaken via logistic regression, using the sequence difference as the outcome with sequence in the model. The log odds ratio derived from this analysis would be the same as (4.1.51) and the test statistic would be (4.1.52). To attain an estimate of the odds-ratio and confidence interval equivalent to the McNemar test one must exponentiate and then square root the $\log(OR_p)$ from the analysis.

Table 4-13. Summary table of period adjusted analysis of hypothetical cross-over trial

	Treatment Difference	
Sequence Difference	-1	+1
A-B	p_{a_1}	p_{b_1}
B-A	p_{a_2}	p_{a_2}
	$\bar{p}_{-1}$	$\bar{p}_{+1}$

If one was to calculate (4.1.52) by hand then one would need to know the variance of log(ψ). However, one knows that logψ=0.5log(OR_p) and that the variance of log(OR_p) is

$$(\text{var}[\log OR_p])^{-1} = \frac{n_d}{12}\left[1 - \sum_{i=-1}^{1} \bar{p}_i^3\right] = (\bar{p}_{+1} + \bar{p}_{-1})^3 - \bar{p}_{+1}^3 - \bar{p}_{-1}^3 = \frac{n}{4}\bar{p}_{+1}(1 - \bar{p}_{+1}), \quad (4.1.53)$$

where, $\bar{p}_i$ is the mean proportion anticipated in category i, for example, $\bar{p}_1 = (p_{a1} + p_{b1})/2$ (where $p_{a1} = a_1 / n_{AB}$ and $p_{b1} = b_1 / n_{AB}$) as described in Table 4.13. Thus, one can combine (4.1.51) with (4.1.53) to obtain the test statistic (4.1.52). If there is no period effect, then (4.1.48) can be rewritten as

$$\frac{n_d}{4}\frac{1}{2}\left(\frac{a_1}{n_{AB}} + \frac{b_2}{n_{BA}}\right)\left(\frac{b_1}{n_{AB}} + \frac{a_2}{n_{BA}}\right)\frac{1}{2} = \frac{n_d}{4}\frac{1}{2}\left(\frac{n_{10} + n_{01}}{n_d}\right)\left(\frac{n_{10} + n_{01}}{n_d}\right)\frac{1}{2} = \frac{(n_{10} + n_{01})^2}{16 n_d},$$

and as var(logψ)=var(0.5logOR_p) and ($n_{10} + n_{01}$)=n, one attains

$$\text{var}\left(\frac{\log OR}{2}\right) = \left(\frac{4}{n_{10} + n_{01}}\right).$$

Additionally, for ψ close to 1 ($0.33 \leq \psi \leq 3$), log(ψ) can be approximated by

$$\log(\psi) \approx \frac{2(\psi - 1)}{\psi + 1} = \frac{2(n_{10} - n_{01})}{n_{10} + n_{01}},$$

and thus, the test statistic (4.1.52) can thus be derived

$$\frac{(\log\psi)^2}{\text{var}(\log\psi)} = \frac{4(n_{10} - n_{01})^2}{(n_{10} + n_{01})^2}\cdot\frac{(n_{10} + n_{01})}{4} = \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}} \sim \chi_1^2.$$

Therefore, (4.1.52) is approximately equal to the McNemar test algebraically if there is no period effect. The consequence of this result is that if by default one does a period adjusted analysis, and there was truly no period effect, then one would obtain equivalent results to the McNemar test. This is analogous to the period-adjusted t-test and paired t-test described in Chapter 2.

In summary, if a period adjusted analysis were undertaken on data where there is no period effect then there would be no effect on the inference. The converse though is not true. Imagine there are two treatment sequences AB and BA with the odds-ratios for each treatment sequence defined as

$$\psi_{AB} = \frac{ka_1}{b_1} = k\psi \text{ and } \psi_{BA} = \frac{kb_2}{a_2} = \frac{k}{\psi},$$

where k (k<1.00) is the known period effect. From (3.51) it is therefore evident that for the special case of $a_1 = a_2$ and $b_1 = b_2$ an unbiased estimate of the odds-ratio is obtained

no matter what the value of the odds-ratio and k. However, if the period difference had been ignored with the data simply pooled the data across the sequences the naïve estimate of the odds-ratio, assuming k and ψ are known, is defined as

$$\psi = \frac{\psi k(\psi + k) + \psi(\psi k + 1)}{k(k\psi + 1) + (\psi + k)}. \qquad (4.1.54)$$

The bias estimated from (4.1.54) is given in Table 4.14 for different values of k. It is evident therefore that by ignoring a possible period effect the results are becoming biased towards the null hypothesis, with the bias increasing with increasing effect size (in absolute terms but not relatively). Overall though, with the exception of large period differences, the bias is relatively small.

Table 4-14. Bias in estimated odds-ratio through ignoring possible period effects

	Odds-Ratio						
k	1.00	1.25	1.50	1.75	2.00	3.00	4.00
0.50	1.000	1.220	1.435	1.647	1.857	2.684	3.500
0.60	1.000	1.233	1.463	1.691	1.918	2.818	3.711
0.70	1.000	1.241	1.481	1.721	1.959	2.908	3.853
0.80	1.000	1.247	1.493	1.738	1.984	2.963	3.941
0.90	1.000	1.249	1.498	1.747	1.996	2.992	3.987
1.00	1.000	1.250	1.500	1.750	2.000	3.000	4.000

4.2.2.16. Sample Size Estimation

4.2.2.17. Population Effects Assumed Known

To calculate the discordant sample size, allowing for any possible period difference the following formulas can be used

$$N_d = \frac{6\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 / (\log OR_p)^2}{/\left[1 - \sum_{i=1}^{1} \bar{p}_i^3\right]} = \frac{2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 / (\log OR_p)^2}{\bar{p}_{\cdot 1}(1 - \bar{p}_{\cdot 1})}, \qquad (4.1.55)$$

where OR_p and $\bar{p}_{\cdot 1}$ are as defined in (4.1.51) and (4.1.53) respectively. For known values of ψ and k Table 4.15 can be derived (similar to Table 4.9) and as a consequence the requisite discordant and total sample size calculated.

Table 4-15. Summary table of period adjusted responses expected in hypothetical cross-over trial

	Treatment Difference		
Sequence Difference	-1	1	Total
A-B	$\frac{\psi}{\psi + k}$	$\frac{k}{\psi + k}$	1
B-A	$\frac{1}{\psi k + 1}$	$\frac{\psi k}{\psi k + 1}$	1

Similarly to earlier in the chapter, a table for each sequence, given in Table 4.16, can be derived using the marginal totals expected from a parallel group study with the expected odds ratio (where $\psi = OR$).

Table 4-16. Summary of hypothetical cross-over trial for each treatment sequence

a. Sequence AB

		Treatment B		
		1	0	
Treatment A	1	$\frac{P_A^2}{(1-P_A)ORk + P_A}$	$\frac{(1-P_A)P_A ORk}{(1-P_A)ORk + P_A}$	P_A
	0	$\frac{P_A(1-P_A)}{(1-P_A)ORk + P_A}$	$\frac{(1-P_A)^2 ORk}{(1-P_A)ORk + P_A}$	$1-P_A$
		$\frac{P_A}{(1-P_A)ORk + P_A}$	$\frac{(1-P_A)ORk}{(1-P_A)ORk + P_A}$	1

a. Sequence BA

		Treatment A		
		1	0	
Treatment B	1	$\frac{kP_A^2}{(1-P_A)OR + P_A}$	$\frac{(1-P_A)P_A k}{(1-P_A)OR + kP_A}$	$\frac{kP_A}{(1-P_A)OR + kP_A}$
	0	$\frac{P_A(1-P_A)OR}{(1-P_A)OR + kP_A}$	$\frac{(1-P_A)^2 OR}{(1-P_A)OR + kP_A}$	$\frac{(1-P_A)OR}{(1-P_A)OR + kP_A}$
		P_A	$1-P_A$	1

Table 4.16 can thus be used to estimate the expected proportion discordant so that the total sample size can be estimated using the two-stage approach described earlier. Therefore, using (4.1.55) and Table 4.16, Table 4.17 can be derived for various values of k, p_1 and the odds-ratio. From this table for k=1 it seems that the estimate sample sizes are slightly smaller compared to equivalent data in Table 4.10. For decreasing values of k the sample size is modestly increased.

Table 4-17. Sample size estimates for a cross-over trial for various expected outcome responses for a given treatment (p_A), period effects (k) and odds-ratios for a two sided type I error rate of 5% and 90% power

		Odds-Ratio					
k	p_A	1.25	1.50	1.75	2.00	3.00	4.00
0.70	0.20	2869	911	496	333	143	98
	0.40	1851	570	303	199	82	56
	0.60	1773	527	272	175	67	45
	0.80	2513	718	358	223	78	49
0.80	0.20	2835	901	492	336	142	94
	0.40	1819	560	298	197	81	52
	0.60	1741	518	268	172	66	41
	0.80	2482	709	354	221	77	44
0.90	0.20	2819	897	489	399	142	93
	0.40	1803	556	296	195	80	52
	0.60	1725	513	265	171	66	41
	0.80	2467	705	352	220	77	44
1.00	0.20	2814	895	489	329	142	93
	0.40	1798	554	295	195	80	52
	0.60	1720	512	265	171	66	41
	0.80	2462	704	351	219	77	44

In summary therefore this section has gone, in detail to recommend ignoring the effect of period in the sample size calculations and use (4.1.33).

4.2.2.18. Sensitivity Analysis and Population Effects Assumed Unknown

Following on from the arguments in the previous sub-section the results for assessing sensitivity (4.1.49) and allow for the imprecision in the marginal control estimate (4.1.50) when period adjustment is not being allowed for in the sample size estimation could be extended to period adjusted sample sizes.

4.2.3. Advantages of Cross-over Trials over Parallel Group Designs

As well as the obvious advantage of a well designed cross-over trial potentially requiring the same total sample size as just one arm of a parallel group, i.e. half the total

sample size, cross-over trials also convey another advantage in the estimation of treatment effects as illustrated in the hypothetical example in Table 4.18.

Imagine one had a parallel group trial designed to compare the outcome of two treatments, where the outcome takes a binary form. Suppose too that a known prognostic factor, gender say, exists but that there is perfect balance with respect to this factor and no interaction between the factor and treatment (see Table 4.18a) with an odds-ratio equal to 3 in each sub group. If one collapsed the data down and ignored the covariate the estimated odds-ratio is biased down to 2.78 (Table 4.18c). However, as a cross-over trial assesses an effect within subject the estimate of the treatment effect is not influenced by any between subject factors (if there is no interactions). This is evidenced by Table 4.18b (derived from the marginal totals of Table 4.18a) and Table 4.18d.

Table 4-18. Hypothetical data from a cross-over and parallel group trial

A. Parallel group broken down by gender

Males

		Outcome		
		1	0	Total
Treatment	A	225	75	300
	B	150	150	300
	Total	375	225	600

Females

		Outcome		
		1	0	Total
Treatment	A	150	150	300
	B	75	225	300
	Total	225	375	600

B. Cross over broken down by gender

Males

		Treatment B		
		1	0	Total
Treatment A	1	112.5	112.5	225
	0	37.5	37.5	75
	Total	150	150	300

Females

		Treatment B		
		1	0	Total
Treatment A	1	37.5	112.5	150
	0	37.5	112.5	150
	Total	75	225	300

* - there are fractions of subjects in the table but the table is only for illustration

C. Parallel group overall

		Outcome		
		1	0	Total
Treatment	A	375	225	600
	B	225	375	600
	Total	600	600	1200

D. Cross over overall

		Treatment B		
		1	0	Total
Treatment A	1	150	225	375
	0	75	150	225
	Total	225	375	600

A caveat should be added here in that this comparison is an "all things being equal" one as cross-over trials by their design are prone to biases and problems to which parallel group trials are not.

As an aside this hypothetical example nicely illustrates a fallacy quoted for binary data that adjusting for covariates inflates the sample size through increasing the standard error [Whitehead, 1993; Robinson and Jewell, 1991]. In the example the unadjusted log-odds-ratio does indeed have a smaller standard error of 0.119 compared to 0.125 for the analysis adjusting for gender. A standard error has increased by 5% through covariate adjustment. However, this effect is more than swamped by the bias in the log odds ratio, with the unadjusted log-odds-ratio being 1.022 (odds-ratio=2.78) compared to an adjusted 1.099 (odds-ratio =3.00) which is a 7.5% increase in the log-odds-ratio by adjusting. Thus, bias introduced by not adjusting will always pull the estimate nearer to the unity. This issue will be discussed in detail in section (4.6) of this chapter.

4.3. Non-Inferiority Trials

Going against the ordering of previous chapters, in this chapter, non-inferiority trials will be discussed before equivalence trials. The reason for this is that the issues under pinning non-inferiority trials are more established than for equivalence trials. However, the points raised in this sub section can be generalised to the later discussions on equivalence trials.

Before describing sample size calculations for non-inferiority trials it is pertinent first to recall the definition of the null (H_0) and alternative (H_1) hypotheses:

H_0: A given treatment is inferior with respect to the mean response ($\pi_A \geq \pi_B$).

H_1: The given treatment is non-inferior with respect to the mean response ($\pi_A < \pi_B$).

These hypotheses can be written in terms of a clinical difference, d [CPMP, 2000; Chen, Tsong and Kang, 2000; Chan, 2003]

$H_0: \pi_A - \pi_B \geq d$.

$H_1: \pi_A - \pi_B < d$.

The issue to highlight here is that under both the null and alternative there is a none zero difference between treatments. This issue was first highlighted by Dunnett and Gent [1977] the implications of which will now be discussed in detail.

4.3.1. Parallel Group Trials

4.3.1.1. Sample Size with the Population Effects Assumed Known

4.3.1.2. Proportional Difference

From Chapter 1 one requires the following

$$Var(S) = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{(d - \Delta)^2}. \qquad (4.2.1)$$

and as with superiority Var(S) can be defined as

$$Var(S) = \frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}. \qquad (4.2.2)$$

Where p_A is the proportion of responses expected in n_A subjects on treatment A and p_B is the expected proportion of responses in n_B subjects on treatment B. For the special case of $n_A = n_B$ (4.2.2) can be substituted in to (4.2.1) (replacing Δ with $p_A - p_B$) giving a direct estimate of the sample size

$$n_A = \frac{(p_A(1-p_A)+p_B(1-p_B))(Z_{1-\beta}+Z_{1-\alpha})^2}{((p_A-p_B)-d)^2}. \qquad (4.2.3).$$

This equation is similar in form to that of the result for superiority trials given earlier in this chapter

$$n_A = \frac{(p_A(1-p_A)+p_B(1-p_B))(Z_{1-\beta}+Z_{1-\alpha/2})^2}{((p_A-p_B)-d)^2}. \qquad (4.2.4)$$

Although not discussed in the section on superiority trials (4.2.4) can actually be written as [Machin, Campbell, Fayers et al, 1997]

$$n_A = \frac{\left(Z_{1-\alpha/2}\sqrt{2\bar{p}(1-\bar{p})}+Z_{1-\beta}\sqrt{p_A(1-p_A)+p_B(1-p_B)}\right)^2}{((p_A-p_B)-d)^2}, \qquad (4.2.5)$$

where $\bar{p}=(p_A+p_B)/2$. Note that under the superiority null hypothesis $p_A = p_B$. Hence, $Z_{1-\alpha/2}$ is multiplied by the variance under the null hypothesis and $Z_{1-\beta}$ is multiplied by the variance under the alternative hypothesis i.e. (4.2.5) can be expressed as

$$n_A = \frac{\left(Z_{1-\alpha/2}\sqrt{\text{Variance under Null}}+Z_{1-\beta}\sqrt{\text{Variance under the Alternative}}\right)^2}{((p_A-p_B)-d)^2}. \qquad (4.2.6)$$

Practically, however, use is made of the following result for superiority trials, $p_A(1-p_A)+p_B(1-p_B) \approx 2\bar{p}(1-\bar{p})$, which as evidence empirically from Table 4.19 (text in bold highlights differences) holds as a reasonable approximation hence enabling (4.2.4) to be used.

Table 4-19. Variances estimated from two different results for different expected treatment responses p_A and p_B

a. $p_A(1-p_A)+p_B(1-p_B)$

	p_B								
p_A	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.10	0.18	0.25	0.30	0.33	**0.34**	**0.33**	**0.30**	**0.25**	**0.18**
0.20	0.25	0.32	0.37	0.40	0.41	**0.40**	**0.37**	**0.32**	**0.25**
0.30	0.30	0.37	0.42	0.45	0.46	0.45	**0.42**	**0.37**	**0.30**
0.40	0.33	0.40	0.45	0.48	0.49	0.48	0.45	**0.40**	**0.33**
0.50	**0.34**	0.41	0.46	0.49	0.50	0.49	0.46	0.41	**0.34**
0.60	**0.33**	**0.40**	0.45	0.48	0.49	0.48	0.45	0.40	0.33
0.70	**0.30**	**0.37**	**0.42**	0.45	0.46	0.45	0.42	0.37	0.30
0.80	**0.25**	**0.32**	**0.37**	**0.40**	0.41	0.40	0.37	0.32	0.25
0.90	**0.18**	**0.25**	**0.30**	**0.33**	**0.34**	0.33	0.30	0.25	0.18

b. $2\bar{p}(1-\bar{p})$

	p_B								
p_A	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.10	0.18	0.26	0.32	0.38	**0.42**	**0.46**	**0.48**	**0.50**	**0.50**
0.20	0.26	0.32	0.38	0.42	0.46	**0.48**	**0.50**	**0.50**	**0.50**
0.30	0.32	0.38	0.42	0.46	0.48	0.50	**0.50**	**0.50**	**0.48**
0.40	0.38	0.42	0.46	0.48	0.50	0.50	0.50	**0.48**	**0.46**
0.50	**0.42**	0.46	0.48	0.50	0.50	0.50	0.48	0.46	**0.42**
0.60	**0.46**	**0.48**	0.50	0.50	0.50	0.48	0.46	0.42	0.38
0.70	**0.48**	**0.50**	**0.50**	0.50	0.48	0.46	0.42	0.38	0.32
0.80	**0.50**	**0.50**	**0.50**	**0.48**	0.46	0.42	0.38	0.32	0.26
0.90	**0.50**	**0.50**	**0.48**	**0.46**	**0.42**	0.38	0.32	0.26	0.18

Note, the practical application of this result will be used further in the section on trials for a given precision. Now for non-inferiority trials one requires the following result for the estimation of sample sizes.

$$n_A = \frac{\left(Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)}+Z_{1-\beta}\sqrt{p_A(1-p_A)+p_B(1-p_B)}\right)^2}{((p_A-p_B)-d)^2}, \quad (4.2.7)$$

where

$$\frac{\tilde{p}_A(1-\tilde{p}_A)}{n_A}+\frac{\tilde{p}_B(1-\tilde{p}_B)}{n_B}, \quad (4.2.8)$$

is an estimate of the variance under the null hypothesis, which has that $p_A \neq p_B$. As the estimates of p_A and p_B affect the estimate of the variance the definition of the null hypothesis hence influences the variance under this hypothesis. There are a number of ways of estimating (4.2.8) and will be discussed now.

4.3.1.3. Method 1 – Using Anticipated Responses

The first method of estimating the variance under the null hypothesis is simply to replace $\tilde{p}_A$ and $\tilde{p}_B$ with anticipated estimates of the response, p_A and p_B [Dunnett and Gent, 1977; Farrington and Manning, 1990]. Hence, the variance under the null hypothesis becomes

$$\frac{p_A(1-p_A)}{n_A}+\frac{p_B(1-p_B)}{n_B}. \qquad (4.2.9)$$

The sample size can hence be estimated from

$$n_A=\frac{(Z_{1-\beta}+Z_{1-\alpha})^2(p_A(1-p_A)+p_B(1-p_B))}{((p_A-p_B)-d)^2}, \qquad (4.2.10)$$

and the same result as (4.2.3).

4.3.1.4. Method 2 –Using Anticipated Responses in Conjunction with the Non-Inferiority Limit

The second method is to estimate $\tilde{p}_A$ and $\tilde{p}_B$ from [Dunnett and Gent, 1977]

$$\tilde{p}_A=(p_A+p_B+d)/2,$$

$$\tilde{p}_B=(p_A+p_B-d)/2, \qquad (4.2.11)$$

where d is the non-inferiority limit. Hence, (4.2.11) can be applied to the estimate of the variance (4.2.8) and an estimate of the sample size can be obtained from

$$n_A=\frac{\left(Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)}+Z_{1-\beta}\sqrt{p_A(1-p_A)+p_B(1-p_B)}\right)^2}{((p_A-p_B)-d)^2}. \qquad (4.2.12)$$

For (4.2.11) to be used the following inequality must hold [Farrington and Manning, 1990]

$$\max\{-d,d\}<p_A+p_B<2+\min\{-d,d\}. \qquad (4.2.13).$$

Farrington and Manning state that (4.2.13) is "easily violated". Which is true but only if one enters stupid values for d, p_A and p_B i.e. if one set d=0.20 where one expected a response rate of 0.90 for both regiment A and B. Patently this is a nonsensical limit for such a high response rate (for d=0.10, a more sensible limit, (4.2.13) would not be violated). Hence, although "easily violated" (4.2.13) holds for all practical limits of d, p_A and p_B.

4.3.1.5. Method 3 – Using Maximum Likelihood Estimates

The third method is to use maximum likelihood estimates for $\tilde{p}_A$ and $\tilde{p}_B$ [Farrington and Manning, 1990; Miettinen and Nurminen, 1985; Koopman, 1984] defined as

$$\tilde{p}_A = 2u\cos(w) - \frac{b}{3a},$$

$$\tilde{p}_B = \tilde{p}_A + d_1, \qquad (4.2.14)$$

to enter into (4.2.8) where $d_1 = p_A(1-d)d$, $b = -(2 + p_A + p_B - 3d)$, a=2, $v = \left(b^2/27a^3 - bc/6a^2 + d_1/2a\right)$, $w = \left[\pi + \cos^{-1}\left(v/u^3\right)\right]/3$, $u = sign(v)\sqrt{b^2/9a^2 - c/3a}$ and $c = d^2 - 2d(p_A + 1) + p_A + p_B$. With (4.2.14) and (4.2.7) an estimate of the sample size can be estimated from -similar in form to (4.2.12)

$$n_A = \frac{\left(Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A) + \tilde{p}_B(1-\tilde{p}_B)} + Z_{1-\beta}\sqrt{p_A(1-p_A) + p_B(1-p_B)}\right)^2}{((p_A - p_B) - d)^2}. \qquad (4.2.15)$$

4.3.1.6. Comparison of the Three Methods of Sample Size Estimation

As evidenced by their descriptions the three methods for estimating the variances under the null hypothesis are markedly different. As a result they give quite different estimates of the variance and as a consequence different estimates for the sample size.

To compare the three methods of sample size estimation a simulation was undertaken. The simulation was undertaken for different p_A and p_B proportions between 0.70 and 0.90 and non-inferiority limits, d, between 0.05 and 0.20. These values were chosen, as they are quite common responses for anti-microbial non-inferiority studies. For each d, p_A and p_B the sample size was iterated until the required power was reached. 100,000 simulations were undertaken to estimate the power for each n, d, p_A and p_B. The simulation was undertaken in SAS [1990].

The simulation was repeated for 4 different methods of calculating confidence intervals: Normal approximation; Normal approximation with continuity correction; Wilson's score method and Wilson's score method with continuity correction. The different methods of calculating the confidence intervals were described in Chapter 3. The simulations were stopped when the requisite proportion of simulations (90%) had an upper tail of a 95% confidence interval less than d. The Normal approximation with continuity correction gave substantially larger estimates of the sample size compared to the other 3 and are not included here.

Table 4-20. Sample sizes for a non-inferiority study estimated through 3 alternative methods for 90% power and a type I error rate of 2.5%

			Sample Size Method		
p_A	p_A-p_B	Limit	Method 1	Method 2	Method 3
0.80	+0.10	0.15	1556	1540	1534
	+0.05	0.10	1466	1452	1471
		0.15	366	359	375
	0.00	0.05	1346	1342	1369
		0.10	337	334	352
		0.15	150	147	162
	+0.05	0.05	303	303	318
		0.10	135	134	149
		0.15	76	74	88
	+0.10	0.05	117	118	130
		0.10	66	66	79
		0.15	43	42	54
0.85	+0.10	0.15	1324	1309	1263
	+0.05	0.10	1209	1199	1209
		0.15	303	296	312
	0.00	0.05	1072	1069	1099
		0.10	268	265	287
		0.15	120	116	135
	-0.05	0.05	229	229	249
		0.10	102	101	120
		0.15	58	56	73
0.90	+0.05	0.10	915	905	896
	0.00	0.05	757	754	791
		0.10	190	186	215

Table 4.20 gives the sample sizes for the different methods of sample size estimation and Table 4.21 gives summary statistics for the ratio of the estimated sample size of over the simulations.

Within the parameters of the simulation it seems that which sample size method to use depends on preference. Method 1 gives the closest estimates compared to simulations. Method 3 one can at least almost guarantee one would never be underpowered. From now on method 1 is the method that will be described.

Table 4-21. Summary statistics comparing the different methods of sample size estimation for a non-inferiority study through simulation and three alternative methods (ratio of calculated to simulation)

		Method of Sample Size Calculation		
CI Calculation	Statistic	Method 1	Method 2	Method 3
Normal Approximation	Minimum	0.93	0.90	0.95
	Quartile 1	0.99	0.97	1.02
	Median	1.00	0.99	1.05
	Quartile 3	1.00	1.00	1.15
	Maximum	1.01	1.05	1.30
Wilson	Minimum	0.82	0.81	0.96
	Quartile 1	0.97	0.97	1.03
	Median	1.00	0.98	1.05
	Quartile 3	1.00	0.99	1.11
	Maximum	1.02	1.03	1.28
Wilson with Continuity Correction	Minimum	0.82	0.81	0.95
	Quartile 1	0.97	0.95	1.02
	Median	0.99	0.98	1.05
	Quartile 3	1.00	1.00	1.11
	Maximum	1.01	1.05	1.28

4.3.1.7. Odds-ratio

Remember the following result for non-inferiority studies the variance of the measure of effect must satisfy

$$Var\,(S) = \frac{(d-\Delta)^2}{(Z_{1-\alpha} + Z_{1-\beta})^2}, \qquad (4.2.16)$$

and that the variance about the log-odds-ratio can be approximated by [Whitehead, 1993]

$$Var\,(S) = \frac{6}{n\left(1 - \sum_{i=1}^{2} \bar{p}_i^3\right)}, \qquad (4.2.17)$$

where $\bar{p}_i$ is the average response each outcome category ($\bar{p}_1 = (p_A + p_B)/2$ and $\bar{p}_2 = 1 - \bar{p}_1$). By equating (4.1.7) with (4.1.6) one requires

$$n = \frac{6\left[Z_{1-\beta} + Z_{1-\alpha}\right]^2}{\left[1 - \sum_{i=1}^{2} \bar{p}_i^3\right](\log(OR) - d)^2}, \qquad (4.2.18)$$

where d in this instance is the non-inferiority limit on the log OR scale. In Chapter 3 appropriate values were quoted as either log(0.43), log(0.50) or log(0.55) as well as that of log(0.47).

Table 4-22. Sample sizes for different non-inferiority limits on the odds-ratio scale and anticipated responses for 90% power and type I error of 2.5%

	Odds-	Limit of 0.43		Limit of 0.47		Limit of 0.50		Limit of 0.55	
p_A	Ratio	Formula	Simulation	Formula	Simulation	Formula	Simulation	Formula	Simulation
0.80	0.70	498	508	745	762	1044	1061	2031	2066
	0.80	319	330	435	445	557	567	876	893
	0.90	234	244	302	307	369	373	525	530
	1.00	185	189	231	235	274	284	368	372
	1.10	154	158	187	192	218	228	282	292
	1.20	132	137	158	163	181	186	228	233
	1.40	105	109	122	127	137	142	167	171
0.85	0.70	612	629	915	933	1282	1317	2496	2530
	0.80	396	406	539	549	690	700	1085	1103
	0.90	292	302	377	387	460	471	655	666
	1.00	232	242	290	300	344	348	462	466
	1.10	194	204	236	247	275	279	355	360
	1.20	167	172	200	205	230	240	289	299
	1.40	133	138	156	160	175	179	212	217
0.90	0.70	848	873	1268	129	1778	1822	3460	3527
	0.80	553	571	753	771	965	982	1518	1543
	0.90	411	421	531	548	648	653	923	940
	1.00	328	339	410	421	487	497	654	664
	1.10	275	286	336	340	391	401	505	516
	1.20	239	249	286	296	328	338	412	417
	1.40	191	201	223	234	251	261	304	315

It was on the odds-ratio scale that Dunnett and Gent recommended inference be based [Dunnett and Gent, 1977]. Not least, as discussed in Chapter 3, because there is a sufficient statistic for the odds-ratio, which leads to exact confidence intervals to be calculable.

As with assessing non-inferiority on the proportional scale (4.1.18) can be rewritten to account for the variability under the null and alternative hypothesis. This is because the variance is estimated from p_A and p_B, which, as discussed earlier in this chapter, differ according to the null and alternative hypothesis. However, (4.1.18) gives the most conservative estimate of the variance as approaches such as method 3 (if applied to the anticipated proportions used to estimate the variance) would reduce the variance estimate on the log-odds scale - the opposite to that for the proportional scale (see Chapter 3).

A simulation was undertaken to assess (4.1.18) for different p_A and p_B proportions between 0.70 and 0.90 and non-inferiority limits, d, log(0.43), log(0.47), log(0.50) and

log(0.55). For each d, p_A and p_B the sample size was iterated until the required power was reached. 100,000 simulations were undertaken to estimate the power for each n, d, p_A and p_B. The simulation was undertaken in SAS [1990]. The simulations were stopped when the requisite proportion of simulations (90%) had a lower tail of the 95% confidence interval greater than d.

Table 4.22 gives the sample sizes calculated from (4.1.18) along with the equivalent sample sizes estimated from simulation. It seems from this table that (4.1.18) underestimates the sample size a little compared to the simulations.

4.3.1.8. Proportional Difference Versus Odds-Ratios

The comparison of the odds-ratio and absolute risk is not straightforward for non-inferiority trials. As discussed in Chapter 3, when setting a non-inferiority limit on the odds-ratio scale one has the prime advantage of a constant margin, which will vary, in terms of the absolute difference, depending on the overall response anticipated. This is opposed to working with the absolute difference where a stepped margin may need to be applied. The issue is debatable however. For modeling purposes the log-odds scale is preferable – not least as it allows for an adjustment for covariates. However for decision analysis, the probability scale is probably the most relevant.

Table 4-23. Comparison of sample sizes calculated on the odds-ratio and proportional scale - assuming $p_A=p_B$

Anticipated Response Rate	Margin	Proportional	Sample Size: Non-Inferiority Limit for the Odds-Ratio 0.43	0.47	0.50	0.55
90%	-10%	190	328	410	487	654
85%	-15%	120	232	290	344	462
80%	-15%	150	185	231	274	368
75%	-20%	99	158	197	234	314
70%	-20%	111	141	176	209	281

Table 4.23 gives the sample size required for different anticipated response rates using calculations based on the absolute difference, from (4.2.10), and the odds-ratio, from (4.2.18). For the odds-ratio different non-inferiority margins were used.

Table 4.23 highlights one disadvantage of using the odds-ratio in that it consistently returns a greater sample size than that for the "equivalent" proportional difference calculation.

Note equivalent was put in speech marks, as the exact equivalent odds-ratio to that of the proportional difference is not being used. However, by the same token the table highlights the disadvantage to working on the absolute scale. Whilst the odds-ratio

moves smoothly up in terms of sample size (requiring the smallest sample size for a response of 0.70 and the greatest for a response of 0.90) for the proportional difference the sample size does not. With the fixed margin of 0.20, an anticipated response rate of 75% would require a sample size less than that of 70%. However, there is a subsequent step up when the response reaches 0.80 and a tighter margin. Working on the absolute risk scale therefore leaves one's calculations very sensitive to assumptions about the anticipated response rates.

4.3.1.9. Worked Example

An investigator wishes to design a trial where the anticipated response rate on the active control is 85%. The investigator also expects an 85% response rate on the investigative therapy. Using an odds-ratio of 0.50 for the non-inferiority limit Table 4.22 gives the sample size as being 344 patients per arm.

In comparison working on the proportional scale, with the same anticipated responses, but with a non-inferiority limit of 15% one requires just 120 patients per arm.

Note here that although the sample sizes here seem quite disparate for the odds-ratio scale compared to the proportional scale, one must bear in mind that one is not comparing like with like. For an anticipated control response of 85%, an odds-ratio of 0.5 equates to a 11.1% difference a little short of 15%.

4.3.1.10. Sensitivity Analysis About the Estimates of the Population Effects Used in the Sample Size Calculations

As highlighted earlier in this chapter for superiority trials, it is the response rate on control, p_A say, to which the study design is sensitive. This response rate in turn feeds into the estimate of variance used in the calculations and, for the absolute difference, the non-inferiority margin used.

As with superiority trials the sensitivity of the non-inferiority study design to the control response rate can be investigated through construction of a 95% confidence interval. The power could then be assessed at the two tails of the confidence interval.

The following result for the absolute difference could be used to investigate the sensitivity of a study. Note that in this formula the study design will be sensitive to both the control response rate (as the non-inferiority margin would change) and the variance,

$$1-\beta = \Phi\left(\sqrt{\frac{n((p_A - p_B)-d)^2}{(p_A(1-p_A)+p_B(1-p_B))}} - Z_{1-\alpha}\right). \qquad (4.2.19)$$

The equivalent formula to investigate the sensitivity of study about an odds-ratio is

$$1-\beta=\Phi\left(\sqrt{n(\log(OR)-d)^2\left[1-\sum_{i=1}^{2}\bar{p}_i^3\right]/6}-Z_{1-\alpha}\right). \qquad (4.2.20)$$

4.3.1.11. Worked Example

Suppose that in the worked example given earlier, the control response rate was assessed from a previous study in 100 patients. It is assumed that the investigative response rate is correct at 85%. Using the Wilson's score method for calculation, the confidence interval indicates that a plausible range for the control response to be between 76.7% and 90.7%.

Table 4-24. Sensitivity analysis for non-inferiority worked example

a. Odds-ratio scale

		95% Confidence Interval	
	Observed	Lower	Upper
Control Response	0.85	0.767	0.907
Investigative Response	0.85	0.850	0.850
Power	90%	100.0%	9.4%

b. Proportional scale

		95% Confidence Interval	
	Observed	Lower	Upper
Control Response	0.85	0.767	0.907
Non-inferiority Margin	0.15	0.15	0.10
Investigative Response	0.90	0.90	0.90
Power	90%	99.7%	17.9%

Table 4.24 gives a breakdown of the sensitivity of the study design to the estimate of the control response rate. As one can see from this table it is the upper point of the confidence interval to which the study is sensitive. The calculations about the odds ratio are more sensitive to this tail (power reduced to 9.4%) than the absolute difference (power reduced to 17.9%).

Note that when undertaking the sensitivity analysis here one is simultaneously assessing the sensitivity of the study to assumptions both about the anticipated variability ($(p_A(1-p_A)+p_B(1-p_B))$ and the mean difference between treatments (p_A-p_B).

Note also that in assessing the sensitivity on the absolute difference scale it was assumed that if one observed a lower than expected control response rate than the original non-inferiority would still be used. However, if a higher than expected response rate was observed that a tighter limit would be used.

4.3.1.12. Calculations Taking Account of the Imprecision of the Estimates of the Population Effects Used in the Sample Size Calculations

As described earlier in this chapter for superiority trials, using appropriate confidence interval methodology around the control response rate, p_A, the power, and, hence iteratively, the sample size can be calculated using numerical methods. By extending this methodology the sample size for a non-inferiority trial, where the proportional difference is of interest, can be estimated from the following result.

$$1-\beta = \frac{1}{0.998}\sum_{pere=0.001}^{0.998} 0.5\left[\Phi\left(\frac{n((p_A - p_B) - d)^2}{\sqrt{(p_{(pere,0.001)}(1-p_{(pere,0.001)}) + p_B(1-p_B))}} - Z_{1-\alpha}\right) + \Phi\left(\frac{n((p_A - p_B) - d)^2}{\sqrt{(p_{(pere,0.001)}(1-p_{(pere,0.001)}) + p_B(1-p_B))}} - Z_{1-\alpha}\right)\right]. \quad (4.2.21)$$

The equivalent calculation for a non-inferiority study design around the odds-ratio would be estimation from

$$1-\beta = \frac{1}{0.998}\sum_{pere=0.001}^{0.998} 0.5\left[\Phi\left(\sqrt{n(\log(OR) - d)^2\left[1 - \sum_{i=1}^{2}\bar{p}_{pere}^3\right]/6} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{n(\log(OR) - d)^2\left[1 - \sum_{i=1}^{2}\bar{p}_{(pere,0.001)}^3\right]/6} - Z_{1-\alpha}\right)\right]. \quad (4.2.22)$$

Tables, which give sample sizes using (4.2.22) and (4.2.21), are given in Tables 4.25 and 4.26 respectively.

This calculation could be questioned for (4.2.21) as here one is assuming that the assumed proportional difference is remaining constant but that the variance is assessed imprecisely. However, both in this instance depend on p_A and as discussed previously in this chapter (and in Chapter 3) the non-inferiority limit also varies depending on p_A (when working on the proportional scale).

Intuitively the calculations are more robust for (4.2.22) as a fixed odds-ratio varies on the absolute scale depending on the anticipated control response rate. Hence, p_A could be considered to only affect the variance.

Table 4-25. Sample sizes for a non-inferiority study, limit of 0.50, on the odds-ratio scale for different precisions around the variance and different anticipated responses for 90% power and type I error of 2.5%

Control Rate	Odds-Ratio	Degrees of Freedom 10	20	30	40	50	100
0.80	0.70	1344	1177	1130	1108	1095	1069
	0.80	724	631	605	592	585	571
	0.90	484	420	402	393	388	379
	1.00	362	313	299	293	289	281
	1.10	289	249	238	233	230	224
	1.20	242	208	199	194	192	187
	1.40	185	158	151	147	145	141
0.85	0.70	1945	1541	1445	1402	1377	1329
	0.80	1060	834	781	757	743	716
	0.90	715	559	523	506	496	478
	1.00	538	419	391	378	371	357
	1.10	433	336	313	303	297	286
	1.20	364	282	262	254	249	239
	1.40	279	215	200	193	189	182
0.90	0.70	4583	2487	2193	2073	2008	1888
	0.80	2523	1360	1196	1130	1093	1026
	0.90	1713	919	807	761	736	690
	1.00	1297	693	608	573	554	519
	1.10	1047	558	489	461	445	417
	1.20	883	470	411	387	374	350
	1.40	681	361	315	297	287	268

Table 4-26. Sample sizes for a non-inferiority study on the absolute difference scale for different precisions around the variance and different anticipated responses for 90% power and type I error of 2.5%

			Control Degrees of Freedom					
Rate	Diff	Limit	10	20	30	40	50	100
0.80	-0.10	0.15	1561	1565	1564	1563	1562	1559
	-0.05	0.10	1495	1485	1479	1475	1473	1468
		0.15	374	372	370	369	369	367
	0.00	0.10	354	347	344	342	341	339
		0.15	158	155	153	152	152	151
	0.05	0.10	147	142	140	139	138	136
		0.15	83	80	79	78	78	77
	0.10	0.10	100	84	75	71	69	67
		0.15	57	51	47	45	44	43
0.85	-0.10	0.15	1399	1372	1360	1353	1348	1337
	-0.05	0.10	1320	1275	1256	1246	1239	1225
		0.15	330	319	314	312	310	307
	0.00	0.10	308	291	284	281	278	274
		0.15	137	130	127	125	124	122
	0.05	0.10	128	116	112	109	108	105
		0.15	71	65	63	62	61	59
0.90	-0.05	0.10	1118	1030	996	977	966	941
	0.00	0.10	255	227	216	210	206	198

4.3.1.13. Worked Example

Suppose that the investigator wishes to redo the calculation from the worked example given earlier to allow for the fact that the control response rate was estimated from 100 patients.

For the same non-inferiority limit of OR=0.5 as previously the sample size should be increased to 357 patients per arm around a 4% increase in the sample size.

Repeating the sample calculations on the absolute difference scale decreases the same size to 122 patients per arm. This is 2 more than the calculation earlier.

4.3.1.14. Calculations Taking Account of the Imprecision of the Estimates Used in the Calculation of Sample Sizes – Bayesian Methods

The percentiles for a posterior control response can be calculated as described in 4.1.1.16. From these percentiles (4.2.21) and (4.2.22) could be used to estimate the sample size allowing for the imprecision in the estimate of the control response rate [Julious, 2004d].

It is best to highlight the points through worked example.

4.3.1.15. Worked Example

For the absolute difference scale with a non-informative prior the sample size is estimate to be 122 patients per arm, which is the same as calculated before.

With a more pessimistic prior, the most likely response being 80% with 90% certainty that it is greater than 75% (from (4.1.26) estimates of a_0 and b_0 of 27.326 and 106.304 are obtained), the sample size estimate is increased to 139 patients per arm.

With a prior that the control response rate observed is about right, the most likely response being 85% with 90% certainty it is greater than 80% (from (4.1.26) estimates of a_0 and b_0 of 18.244 and 98.716 are obtained), the sample size estimate is to 122 patients per arm – the same as for a non-informative prior.

Similar calculations could be done for the odds-ratio.

4.3.1.16. Calculations Taking Account of the Imprecision of the Estimates of the Population Effects with Respect to the Assumptions about the Mean Difference and the Variance Used in the Sample Size Calculations

When one is designing a non-inferiority study, as discussed in Chapter 2 for Normal data, the imprecision in the mean difference as well as the variance may be of importance. This is particularly so for non-inferiority studies (and equivalence studies

described later in the chapter) where the mean response, assessed by p_A, feeds into the assumptions both about the mean difference and the variance.

To allow for the imprecision in the mean difference and variance one could use numerical methods to calculate the sample size on the absolute difference scale and the following result

$$1-\beta = \frac{1}{0.998}\sum_{perc=0.001}^{0.998} 0.5\left[\Phi\left(\frac{n((p_{perc_A}-p_B)-d)^2}{\sqrt{(p_{perc_A}(1-p_{perc_A})+p_B(1-p_B))}}-Z_{1-\alpha}\right)+\Phi\left(\frac{n((p_{perc_A}-p_B)-d)^2}{\sqrt{(p_{perc_A+0.001}(1-p_{perc_A+0.001})+p_B(1-p_B))}}-Z_{1-\alpha}\right)\right]. \quad (4.2.23)$$

Note that in this instance, in contrast to non-inferiority calculations given earlier, a number of issues need to be additionally considered

1. If any of the percentile values around the control response crosses a step (given in Table 4.23) then the non-inferiority margin, d, should be altered accordingly.

2. The investigative response rate, p_B, remains assumed fixed calculated from the initial p_A but not from individual p_{perc_A}.

3. Following on from 2. for instances where $p_{perc_A} - p_B$ exceeds the non-inferiority bound then the power for this percentile (to be averaged across for power calculation) is set to 0.

The equivalent calculation for a non-inferiority study designed around the odds-ratio would be

$$1-\beta = \frac{1}{0.998}\sum_{perc=0.001}^{0.998} 0.5\left[\Phi\left(\sqrt{n}(\log OR_{perc}-d)^2\left[1-\sum_{i=1}^{2}\bar{p}^3_{perc}\right]/6-Z_{1-\alpha}\right)+\Phi\left(\sqrt{n}(\log OR_{perc}-d)^2\left[1-\sum_{i=1}^{2}\bar{p}^3_{(perc+0.001)}\right]/6-Z_{1-\alpha}\right)\right]. \quad (4.2.24)$$

As the odds-ratio does not suffer from the issues of stepped non-inferiority bounds the calculations are relatively more straightforward. However, the following two points should be considered similar to the proportional difference

1. The investigative response rate, p_B, remains assumed fixed calculated from the initial p_A but not from individual p_{perc_A}.

2. Following on from 2. for instances where $OR_{perc} = (p_{perc_A}(1-p_B))/(p_B(1-p_{perc_A}))$ exceeds the non-inferiority bound then the power for this percentile (to be averaged across for power calculation) is set to 0.

Tables 4.27 and 4.28 give sample size calculations using (4.2.24) and (4.2.23) respectively. A couple of points are worth noting from these tables.

Table 4-27. Sample sizes for a non-inferiority study, with a limit of 0.5, on the odds-ratio scale for different precisions around the anticipated control response rate and variance for 90% power and type I error of 2.5%

Control Rate	Odds-Ratio	Degrees of Freedom 50	100	250	500
0.80	0.90	2971	861	504	430
	1.00	1070	500	344	306
	1.10	603	348	261	238
	1.20	411	266	211	195
	1.40	250	183	154	145
0.85	0.90	10717	1424	689	560
	1.00	2324	757	461	397
	1.10	1103	506	347	308
	1.20	691	379	278	253
	1.40	388	255	202	188
0.90	0.90	37602	3993	1177	863
	1.00	19176	1653	750	601
	1.10	4325	988	550	462
	1.20	2033	697	435	377
	1.40	886	441	312	279

Table 4-28. Sample sizes for a non-inferiority study on the absolute difference scale for different precisions around the anticipated control response rate and variance for 90% power and type I error of 2.5%

Control Rate	Pa-Pb	Limit	Degrees of Freedom 50	100	250	500
0.80	0.00	0.15	215	172	148	148
	+0.05	0.15	86	76	75	75
	+0.10	0.15	43	41	41	41
0.85	0.00	0.15	283	150	126	122
	+0.05	0.15	80	64	59	58
0.90	+0.00	0.10	263	214	188	188

The first is that these calculations are only possible if one is optimistic in one's assumptions around the investigative response rate i.e. one should assume the response is equal to or better than the control response rate.

The second is that the odds-ratio for these calculations give very large sample size estimates. In fact these calculations seem to demonstrate that just because you can do the calculation it does not mean you should. Indeed as on the OR scale the margin changes as the control response changes (in terms of absolute differences) one could seriously question the need for these calculations. The recommendation therefore would be simply to use (4.2.22) and Table 4.25 for odds-ratio calculations.

For the absolute difference scale the calculations seem more plausible. This is because the assumed absolute difference (and non-inferiority margin) now changes depending

on the imprecision of the control response. The sample sizes are quite plausible and are in the same region as Table 4.25. The recommendation would therefore be to use (4.2.24) and Table 4.28 when working on the absolute difference scale.

4.3.1.17. Proportional Difference Versus Odds-Ratios - Revisited

It seems at first that the calculations on the odds-ratio scale are overly sensitive, compared to the absolute difference, to assumptions around the variance. In fact this is a function of the properties of the odds-ratio – the fact that a fixed odds-ratio would equate to smaller and smaller differences, as the control response gets greater. In comparison on the absolute difference scale the margins are relatively fixed (all be it stepped) such the same margin could be used, 10%, independent of the anticipated response.

Which statistical analysis and consequent sample size calculation, to use depends on the robustness of one's assumptions. If it is reasonable to have relatively fixed margins then one can work completely on the proportional scale. If one wishes to have more flexible margins then one should work on the odds-ratio scale.

In truth, however, there is no generic answer as to what scale to work on. For example an anticipated response of 90% raises far greater questions (should the margin narrow if a response rate greater than 90% is observed?) than one of 80%. Thus, it is recommended that the decision as to the most calculations be undertaken on a case by case basis with a thorough investigation made as to the sensitivity of one's calculations to the assumptions inherent in them.

4.3.1.18. Worked Example

Revisiting the worked example again where the investigator had a control response rate estimated from a trial with 100 patients.

For the same non-inferiority limit of OR=0.5 as previously the sample size should be increased to 757 patients per arm approximately a 2 fold increase in the sample size compared to the original calculations.

Repeating the sample calculations on the absolute difference scale increases the same size to 150 patients per arm approximately a 20% increase in the sample size.

4.3.1.19. Calculations That Take Account of the Imprecision of the Estimates Effects with Respect to the Assumptions about the Mean Difference and the Variance Used in the Sample Size Calculations – Bayesian Methods.

As previously discussed the percentiles for a posterior control response can be calculated as in 4.1.1.16 and from these percentiles from (4.2.21) and (4.2.22) can be

used to give an estimate of the sample size [Julious, 2004d]. Again it is best to highlight the points through worked example.

4.3.1.20. Worked Example

For the absolute difference scale with a non-informative prior the sample size is estimate to be 152 patients per arm. This is two greater than the sample size calculated before.

With a more pessimistic prior (the most likely response being 90% with 90% certainty that it is greater than 85%), the sample size estimate is increased to 155 patients per arm.

Note that this is more of a pessimistic prior then when just looking at variability as one is now calculating the power for instances when the belief is that the mean difference is in favour of the control treatment. This will adversely affect the sample size.

With a prior that the control response rate observed is about right (the most likely response being 85% with 90% certainty it is greater than 80%) the sample size estimate is increased to 126 patients per arm.

4.3.2. Cross-over Trials

There are a number of papers which deal specifically on the topic of cross-over equivalence trials [Lu and Bean, 1995; Tango 1998, 1999; Nam, 1997; Tang, 2003; Tang, Tang and Chan, 2003]. However, these methodologies are simply extensions of methodologies for superiority cross-over trials and parallel group non-inferiority trials.

Earlier in this chapter it was highlighted how to estimate the sample size for superiority trial one could simply use the sample sizes for parallel group superiority trials and take the sample size per arm to be the total sample size for a cross-over trial. This argument can be extended now to non-inferiority trials. It is therefore recommended to use the parallel group methodologies described in this sub-section of the chapter to estimate the total sample size for a non-inferiority cross-over trial

4.4. As Good as or Better Trials

As discussed in Chapter 2, to calculate the sample size required for an "as good as or better" trial one should apply the methodologies described for superiority and non-inferiority trials.

The big advantage in designing as good as or better trials for binary data is that the non-inferiority limits (as discussed throughout this chapter and Chapter 3) are more clearly

defined. Hence, the criteria for the closed testing procedure are also more straightforward to define.

Other issues with as good as or better trials are either the same as described for Normal data in Chapter 2 or generic and described in Chapter 1. Hence, this chapter will not go into detail on these types of trial now.

One issue that may potentially become more open for debate is the choice of analysis population. As discussed in Chapter 3, for a superiority trials the primary data set is the intention to treat population (ITT); whilst for non-inferiority trials it is both the per protocol data set (PP) and the ITT [CPMP, 2000]. Garrett [2003] has recently challenged this assertion. Highlighting how for binary data there is no bias in the point estimates for the PP population with any conservativeness being down simply to a smaller sample size. However, the regulatory guidance currently advocates the joint primary population approach [CPMP, 2000] and so any sample size calculation should also reflect this.

4.5. Equivalence Trials

For equivalence trials, the null (H_0) and alternative (H_1) hypotheses are defined as:

H_0: A given treatment is inferior with respect to the mean response ($\pi_A \neq \pi_B$).

H_1: The given treatment is equivalent with respect to the mean response ($\pi_A = \pi_B$).

Formally, these hypotheses can be written in terms of a clinical difference, d [CPMP, 2000]

$H_0: \pi_A - \pi_B \geq d$ or $\pi_A - \pi_B \leq -d$.

$H_1: -d < \pi_A - \pi_B < d$.

The issue to highlight here is that like non-inferiority trials under both the null and alternative there is a non-zero difference between treatments [Dunnett and Gent, 1977]. The implications are similar to those for non-inferiority trials discussed earlier and will now be discussed.

4.5.1. Parallel Group Trials

4.5.1.1. Sample Sizes with the Population Effects Assumed Known

4.5.1.2. General Case

4.5.1.3. Proportional Difference

Recall from Chapter 1 that the total Type II error (defined as $\beta = \beta_1 + \beta_2$) is derived from the following result.

$$Z_{1-\beta_1} = \frac{-d-\Delta}{\sqrt{Var(S)}} - Z_{1-\alpha} \text{ and } Z_{1-\beta_2} = \frac{d-\Delta}{\sqrt{Var(S)}} - Z_{1-\alpha}.$$

As discussed in Chapter 2, for equivalence trials for the general case where the expected true mean difference is not fixed to be zero the sample size cannot be derived directly as the total Type II error is the sum of the Type II errors associated with each one-tailed test.

As with non-inferiority trials discussed earlier in this chapter there are a number of approaches for the derivation of the variance under the null ($\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)$) and alternative ($p_A(1-p_A)+p_B(1-p_B)$) hypothesis. The generic solution to estimation the power for a given sample size is thus

$$1-\beta = \Phi\left(\sqrt{\frac{n((p_A-p_B)-d)^2}{p_A(1-p_A)+p_B(1-p_B)}} - \frac{Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)}}{\sqrt{p_A(1-p_A)+p_B(1-p_B)}}\right) + \Phi\left(\sqrt{\frac{n((p_A-p_B)+d)^2}{p_A(1-p_A)+p_B(1-p_B)}} - \frac{Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)}}{\sqrt{p_A(1-p_A)+p_B(1-p_B)}}\right) - 1 \quad (4.3.1)$$

This chapter will now discuss the different methods for estimation the variances.

4.5.1.4. Method 1 – Using Anticipated Responses

As with non-inferiority trials the first method of estimating the variance under the null hypothesis is simply to replace $\tilde{p}_A$ and $\tilde{p}_B$ with anticipated estimates of the response, p_A and p_B [Dunnett and Gent, 1977; Farrington and Manning, 1990]. Hence, the variance under the becomes

$$\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}, \quad (4.3.2)$$

and the power for a given sample size can hence be estimated from

$$1-\beta = \Phi\left(\sqrt{\frac{n_A((p_A-p_B)-d)^2}{p_A(1-p_A)+p_B(1-p_B)}} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{\frac{n_A((p_A-p_B)+d)^2}{p_A(1-p_A)+p_B(1-p_B)}} - Z_{1-\alpha}\right) - 1. \quad (4.3.3)$$

To estimate the sample size one iterates (4.3.3) on the sample size until the nominal power is reached.

4.5.1.5. Method 2 – Using Anticipated Responses in Conjunction with the Equivalence Limit.

The second method is to estimate $\tilde{p}_A$ and $\tilde{p}_B$ from [Dunnett and Gent, 1977]

$$\tilde{p}_A = (p_A + p_B + d)/2,$$

$$\tilde{p}_B = (p_A + p_B - d)/2, \qquad (4.3.4)$$

where d are symmetric equivalence limits. Applying, (4.3.4) to (4.3.1), an estimate of the power for a given sample size can be obtained from

$$1-\beta = \Phi\left(\sqrt{\frac{n((p_A-p_B)-d)^2}{p_A(1-p_A)+p_B(1-p_B)}} - \frac{Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)}}{\sqrt{p_A(1-p_A)+p_B(1-p_B)}}\right) + \Phi\left(\sqrt{\frac{n((p_A-p_B)+d)^2}{p_A(1-p_A)+p_B(1-p_B)}} - \frac{Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)}}{\sqrt{p_A(1-p_A)+p_B(1-p_B)}}\right) - 1 \qquad (4.3.5)$$

One uses this result to iterate to find the required sample size. As for non-inferiority trials to use (4.3.5) the following inequality must hold [Farrington and Manning, 1990]

$$\max\{-d, d\} < p_A + p_B < 2 + \min\{-d, d\}.$$

4.5.1.6. Method 3 – Using Maximum Likelihood Estimates

The third method, again like for non-inferiority trials is to use maximum likelihood estimates for $\tilde{p}_A$ and $\tilde{p}_B$ [Farrington and Manning, 1990; Miettinen and Nurminen, 1985; Koopman, 1984] defined as

$$\tilde{p}_A = 2u\cos(w) - \frac{b}{3a},$$

$$\tilde{p}_B = \tilde{p}_A + d_1,$$

to enter into (4.3.1) where $d_1 = p_A(1-d)d$, $c = d^2 - 2d(p_A+1) + p_A + p_B$, $b = -(2 + p_A + p_B - 3d)$, $w = \left[\pi + \cos^{-1}\left(v/u^3\right)\right]/3$, $u = sign(v)\sqrt{b^2/9a^2 - c/3a}$,a=2, and $v = b^2/27a^3 - bc/6a^2 + d_1/2a$. Hence, an estimate of the sample size can be estimated from

$$1-\beta=\Phi\left(\sqrt{\frac{n((p_A-p_B)-d)^2}{p_A(1-p_A)+p_B(1-p_B)}}-\frac{Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)}}{\sqrt{p_A(1-p_A)+p_B(1-p_B)}}\right) \quad (4.3.6)$$
$$+\Phi\left(\sqrt{\frac{n((p_A-p_B)+d)^2}{p_A(1-p_A)+p_B(1-p_B)}}-\frac{Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)}}{\sqrt{p_A(1-p_A)+p_B(1-p_B)}}\right)-1$$

Table 4-29. Sample sizes for an equivalence study estimated by and 3 alternative methods for 90% power and a type I error rate of 2.5%

			Sample Size Method		
p_A	p_A-p_B	Limit	Method 1	Method 2	Method 3
0.80	-0.10	0.15	1556	1540	1534
	-0.05	0.10	1461	1451	1471
		0.15	366	359	375
	0.00	0.05	1664	1660	1690
		0.10	416	413	433
		0.15	185	182	199
	+0.05	0.10	1209	1199	1334
		0.15	303	296	350
	+0.10	0.15	1051	1035	1330
0.85	-0.10	0.15	1324	1309	1263
	-0.05	0.10	1209	1199	1209
		0.15	303	296	312
	0.00	0.05	1326	1322	1356
		0.10	332	328	353
		0.15	148	144	165
	00.05	0.10	915	905	1077
		0.15	229	223	289
0.90	+0.05	0.10	915	905	895
	0.00	0.05	936	933	973
		0.10	234	231	262

4.5.1.7. Comparison of the Three Methods

A similar comparison to that done earlier in the chapter for non-inferiority trials was undertaken to compare the three methods of sample size estimation through simulation. The simulation was undertaken for different p_A and p_B (between 0.70 and 0.90) and non-inferiority limits, d (between 0.05 and 0.20). For each d, p_A and p_B the sample size was iterated until the required power was reached. 100,000 simulations were undertaken to estimate the power for each n, d, p_A and p_B. The simulation was undertaken in SAS [1990].

Note though that for equivalence trials there is a greater limitation on the calculations as p_A - p_B cannot now exceed either -d or d. For non-inferiority trials there is just one bound.

As with non-inferiority trials earlier the simulation was repeated for 4 different methods of calculating confidence intervals: Normal approximation; Normal approximation with continuity correction; Wilson's score method and Wilson's score method with continuity correction. Again though the Normal approximation with continuity correction gave substantially larger estimates of the sample size compared to the other 3 and are not included.

The simulations were stopped when the requisite proportion of simulations (90%) had a 95% confidence interval wholly contained within (-d, d). Table 4.29 gives the sample sizes for the different methods of sample size estimation and Table 4.30 gives summary statistics for the ratio of the estimated sample size of over the simulations.

With the parameters of the simulation for the equivalence trial comparison methods 1 and 2 give closer estimates of the sample size compared to simulations than for non-inferiority trials described earlier in the chapter, with method 1, being the closest to the simulations. Again method 3 seems over-estimate the sample size when compared to the simulated results.

Through the remainder of this section it will be method 1 that will be method described.

Table 4-30. Summary statistics comparing the different methods of sample size estimation for an equivalence study through simulation and 3 alternative methods (ratio of calculated to simulation)

		Method of Sample Size Calculation		
CI Calculation	Statistic	Method 1	Method 2	Method 3
Normal Approximation	Minimum	0.98	0.95	0.95
	Quartile 1	1.00	0.98	1.01
	Median	1.00	0.99	1.04
	Quartile 3	1.00	0.99	1.11
	Maximum	1.01	1.00	1.31
Wilson	Minimum	0.95	0.93	0.95
	Quartile 1	1.00	0.98	1.01
	Median	1.00	0.99	1.06
	Quartile 3	1.01	1.00	1.11
	Maximum	1.03	1.00	1.31
Wilson with Continuity Correction	Minimum	1.00	0.99	0.96
	Quartile 1	1.01	1.00	1.04
	Median	1.02	1.01	1.09
	Quartile 3	1.04	1.02	1.16
	Maximum	1.08	1.05	1.32

4.5.1.8. Odds-Ratio

Remember again that the variance about the log-odds-ratio can be approximated as [Whitehead, 1993]

$$Var\,(S) = \frac{6}{n\left(1 - \sum_{i=1}^{2} \bar{p}_i^{\,3}\right)}, \qquad (4.3.7)$$

where $\bar{p}_i$ is the average response on each outcome category ($\bar{p}_1 = (p_A + p_B)/2$ and $\bar{p}_2 = 1 - \bar{p}_1$). Consequently an estimate of the sample size for a given power can be estimated from

$$1-\beta = \Phi\left(\sqrt{\left[1 - \sum_{i=1}^{2} \bar{p}_i^{\,3}\right](\log(OR) - d)^2 / 6} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{\left[1 - \sum_{i=1}^{2} \bar{p}_i^{\,3}\right](\log(OR) + d)^2 / 6} - Z_{1-\alpha}\right) - 1, \qquad (4.3.8)$$

where d in this instance is the symmetric equivalence limit on the log scale. For non-inferiority trials discussed earlier suggested values for d were given as: log(0.43), log(0.47), log(0.50) or log(0.55). The rational for their use in non-inferiority trials can be generalised to equivalence trials.

Table 4-31. Sample sizes for different equivalence limits on the odds-ratio scale and anticipated responses for 90% power and type I error of 2.5%

	Odds-	Limit of 0.43		Limit of 0.47		Limit of 0.50		Limit of 0.55	
p_A	Ratio	Formula	Simulate	Formula	Simulate	Forumula	Simulate	Formula	Simulate
0.80	0.70	498	509	745	756	1044	106	2031	2057
	0.80	319	330	435	446	557	568	876	881
	0.90	243	254	311	316	377	388	532	543
	1.00	229	240	285	296	339	344	455	466
	1.10	254	265	323	328	391	402	546	557
	1.20	318	323	424	429	532	543	804	815
	1.40	564	582	829	847	1141	116	2124	2142
0.85	0.70	612	623	915	933	1282	130	2496	2531
	0.80	396	407	539	550	690	701	108	1096
	0.90	303	314	388	399	471	482	663	674
	1.00	287	298	358	369	425	436	571	582
	1.10	320	331	407	425	492	503	688	699
	1.20	403	421	536	547	673	684	101	1028
	1.40	717	735	1054	1080	1452	147	2703	2729
0.90	0.70	848	874	1268	1286	1778	1823	3460	3505
	0.80	553	579	754	772	965	983	151	1536
	0.90	427	445	547	565	663	681	934	952
	1.00	406	424	507	525	602	620	808	826
	1.10	455	473	580	598	700	718	979	1005
	1.20	575	593	766	792	962	988	145	1478
	1.40	1030	1075	1514	1559	2085	2141	3883	3939

As with designing non-inferiority trials (4.3.8) can be rewritten to account for the different variabilities under the null and alternative hypothesis. However, as with non-inferiority trials, (4.3.8) gives the most conservative estimate of the variance.

A simulation was undertaken to assess (4.1.8) for different p_A and p_B proportions between 0.70 and 0.90 and symmetric (on the log scale) equivalence limits, d, log(0.43), log(0.47), log(0.50) and log(0.55). For each d, p_A and p_B the sample size was iterated until the required power was reached. 100,000 simulations were undertaken to estimate the power for each n, d, p_A and p_B. The simulation was undertaken in SAS [1991].

The simulations were stopped when the requisite proportion of simulations (90%) had a 95% confidence interval wholly contained within (-d,d).

Table 4.31 gives the sample sizes calculated from (4.1.8) along with the sample sizes estimated from simulation. From this one can see that (4.1.8) gives sample size estimates close to the simulation, although under estimating little, all be it consistently.

4.5.1.9. Proportional Difference Versus Odds-Ratios

The issues raised in comparing the proportional difference and the odds-ratio for non-inferiority trials can be generalised to equivalence trials and so this comparison of odds ratios and proportional differences will not be made here.

4.5.1.10. Special Case of No Treatment Difference

As with equivalence trials discussed for Normal data in Chapter 2 when the assumption is made of no true difference between treatments the calculations are greatly simplified with a direct estimate of the sample size now possible. This sub-section will now briefly discuss these calculations

4.5.1.11. Proportional Difference

The three methods below are as described earlier for the general case.

4.5.1.12. Method 1 – Using Anticipated Responses

For the special case of no anticipated treatment difference the power can be estimated from

$$1-\beta = 2\Phi\left(\sqrt{\frac{n_A d^2}{p_A(1-p_A)+p_B(1-p_B)}} - Z_{1-\alpha}\right) - 1. \qquad (4.3.9)$$

However, as $p_A = p_B$ (4.3.9) can be rewritten as

$$1-\beta = 2\Phi\left(\sqrt{\frac{n_A d^2}{2\bar{p}(1-\bar{p})}} - Z_{1-\alpha}\right) - 1. \qquad (4.3.10)$$

where $\bar{p} = (p_A + p_B)/2$ interpreted in this instance as the anticipated overall response. Equation (4.3.10) can in turn be rewritten to give a direct estimate of the sample size [Machin, Campbell; Fayers et al, 1997]

$$n_A = \frac{2\left(Z_{1-\beta/2} + Z_{1-\alpha}\right)^2 \bar{p}(1-\bar{p})}{d^2}. \qquad (4.3.11)$$

4.5.1.13. Method 2 – Using Anticipated Responses in Conjunction with the Equivalence Limit

Following on from the arguments for method 1 the power is estimated from

$$1-\beta = 2\Phi\left(\sqrt{\frac{nd^2}{\bar{p}(1-\bar{p})}} - \frac{Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)}}{\sqrt{\bar{p}(1-\bar{p})}}\right) - 1, \qquad (4.3.12)$$

where $\tilde{p}_A = \bar{p} + d/2$ and $\tilde{p}_B = \bar{p} - d/2$ and the inequality turn hence now becomes $\max\{-d, d\} < 2\bar{p} < 2 + \min\{-d, d\}$. From (4.3.12) for a direct estimate of the sample size one gets

$$n_A = \frac{\left(Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)} + Z_{1-\beta/2}\sqrt{2\bar{p}(1-\bar{p})}\right)^2}{d^2}. \qquad (4.3.13)$$

4.5.1.14. Method 3 – Using Maximum Likelihood Estimates

For method 3 $\tilde{p}_A$ and $\tilde{p}_B$ are now a little different

$$\tilde{p}_A = 2u\cos(w) - \frac{b}{3a},$$

$$\tilde{p}_B = \tilde{p}_A + d_1,$$

where $d_1 = \bar{p}(1-d)d$, $c = d^2 - 2\left(d(\bar{p}+1)+\bar{p}\right)$, $b = -(2(1+\bar{p}) - 3d)$, a=2, $v = b^2/27a^3 - bc/6a^2 + d_1/2a$, $u = sign(v)\sqrt{b^2/9a^2 - c/3a}$ and $w = \left[\pi + \cos^{-1}\left(v/u^3\right)\right]/3$.

However, given these definitions the formula for the power

$$1-\beta = 2\Phi\left(\sqrt{\frac{nd^2}{\bar{p}(1-\bar{p})}} - \frac{Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)}}{\sqrt{\bar{p}(1-\bar{p})}}\right) - 1, \qquad (4.3.14)$$

and the sample size

$$n_A = \frac{\left(Z_{1-\alpha}\sqrt{\tilde{p}_A(1-\tilde{p}_A)+\tilde{p}_B(1-\tilde{p}_B)}+Z_{1-\beta 2}\sqrt{2\bar{p}(1-\bar{p})}\right)^2}{d^2}, \qquad 4.3.15)$$

take similar forms to (4.3.12) and (4.3.13) respectively.

4.5.1.15. Odds-Ratio

With the assumption of no true difference between treatments (equivalent to OR=1) the power can be estimated from

$$1-\beta = 2\Phi\left(\sqrt{\left[1-\sum_{i=1}^{2}\bar{p}_i^3\right]d^2/6}-Z_{1-\alpha}\right)-1, \qquad (4.3.16)$$

whilst a direct estimate of the sample size can be obtained from

$$n = \frac{6\left[Z_{1-\beta}+Z_{1-\alpha}\right]^2}{\left[1-\sum_{i=1}^{2}\bar{p}_i^3\right]d^2}. \qquad (4.3.17)$$

4.5.1.16. Worked Example

An investigator wishes to design an equivalence trial where the anticipated response rate on the active control is 85%. The investigator also expects a 85% response rate on the investigative therapy. Using an odds-ratio of 0.50 for the symmetric equivalence limit Table 4.31 gives the sample size as being 425 patients per arm.

In comparison, working on the proportional scale, with the same anticipated responses, but with an equivalence limit of 15%, one would require (from Table 4.30) just 148 patients per arm.

4.5.1.17. Sensitivity Analysis About the Estimates of the Population Effects Used in the Sample Size Calculations

As with superiority and non-inferiority trials discussed earlier in this chapter the sensitivity of an equivalence study design to the control response rate can be investigated through construction of a 95% confidence interval. The power could then be used assessed at the two tails of the confidence interval.

This confidence interval could then be used with (4.3.3), for an absolute difference, and (4.3.8), for an odds-ratio to interrogate the sensitivity of the study to the control response rate.

4.5.1.18. Worked Example

Suppose the control response rate was assessed from a previous study in 100 patients and it is assumed that the investigative response rate is fixed at 85%.

Using the Wilson's score method for calculation, confidence interval indicates that a plausible range for the control response to be between 76.7% and 90.7%.

Table 4.32 gives a breakdown of the sensitivity of the study design to the estimate of the control response rate. As is evidenced from this table this equivalence study is sensitive to both the lower and upper points of the confidence interval – as these both bring the point estimate closer to the equivalence boundary.

Table 4-32. Sensitivity analysis for equivalence worked example

a. Odds-ratio scale

		95% Confidence Interval	
	Observed	Lower	Upper
Control Response	0.85	0.767	0.907
Investigative Response	0.85	0.850	0.850
Power	90%	13.6%	10.7%

b. Proportional Scale

		95% Confidence Interval	
	Observed	Lower	Upper
Control Response	0.85	0.767	0.907
Non-inferiority Margin	0.15	0.15	0.10
Investigative Response	0.90	0.90	0.90
Power	90%	31.2%	19.1%

For the odds ratio calculation the lower and upper tails of the confidence interval have powers of 13.6% and 10.7% respectively. Whilst for the absolute difference the lower and upper tails have 31.2 and 19.1% power respectively.

Note that the same assumptions were made here as for non-inferiority trials earlier in this chapter with respect to the simultaneous assessment of the sensitivity of the study to both increases in anticipated variability $(p_A(1-p_A)+p_B(1-p_B))$ and the mean difference between treatments (p_A - p_B). Another consideration is that if a higher than expected response rate was observed then a tighter limit would be used.

4.5.1.19. Calculations Taking Account of the Imprecision of the Estimates of the Populations Effects Used in the Sample Size Calculations

As for non-inferiority and superiority trials, by using appropriate confidence interval methodology around the control response rate, p_A, the power, and hence the sample size can be calculated using numerical methods for equivalence trials. Hence, where the proportional difference is of interest, the sample size can be estimated from the following result,

$$1-\beta = \frac{1}{0.998}\sum_{perc=0.001}^{0.998}\frac{\lambda_1+\lambda_2}{2}, \qquad (4.3.18)$$

where λ_A and λ_B are defined as

$$\lambda_1 = \Phi\left(\sqrt{\frac{n((p_A-p_B)-d)^2}{(p_{(perc-0.001)_+}(1-p_{(perc-0.001)_+})+p_B(1-p_B))}}-Z_{1-\alpha}\right)+\Phi\left(\sqrt{\frac{n((p_A-p_B)-d)^2}{(p_{(perc)_-}(1-p_{(perc)_-})+p_B(1-p_B))}}-Z_{1-\alpha}\right)-1,$$

$$\lambda_2 = \Phi\left(\sqrt{\frac{n((p_A-p_B)-d)^2}{(p_{(perc+0.001)_+}(1-p_{(perc+0.001)_+})+p_B(1-p_B))}}-Z_{1-\alpha}\right)+\Phi\left(\sqrt{\frac{n((p_A-p_B)-d)^2}{(p_{(perc+0.001)_-}(1-p_{(perc+0.001)_-})+p_B(1-p_B))}}-Z_{1-\alpha}\right)-1.$$

The equivalent calculation for an equivalence study designed around the odds-ratio would be,

$$1-\beta = \frac{1}{0.998}\sum_{perc=0.001}^{0.998}\frac{\eta_1+\eta_2}{2}, \qquad (4.3.19)$$

where η_A and η_B are defined as

$$\eta_1 = \Phi\left(\sqrt{n(\log(OR)-d)^2\left[1-\sum_{i=1}^{2}\bar{p}^3_{perc}\right]/6}-Z_{1-\alpha}\right)+\Phi\left(\sqrt{n(\log(OR)+d)^2\left[1-\sum_{i=1}^{2}\bar{p}^3_{perc}\right]/6}-Z_{1-\alpha}\right)-1$$

$$\eta_2 = \Phi\left(\sqrt{n(\log(OR)-d)^2\left[1-\sum_{i=1}^{2}\bar{p}^3_{(perc+0.001)_+}\right]/6}-Z_{1-\alpha}\right)+\Phi\left(\sqrt{n(\log(OR)+d)^2\left[1-\sum_{i=1}^{2}\bar{p}^3_{(perc+0.001)_-}\right]/6}-Z_{1-\alpha}\right)-1$$

Tables 4.34 and 4.33 give sample sizes using (4.3.18) and (4.3.19) respectively. As with non-inferiority trials discussed earlier in this chapter, intuitively the calculations are more robust for (4.3.19) where a fixed odds-ratio varies on the absolute scale depending on anticipated control response rate.

Table 4-33. Sample sizes for an equivalence study, limit of 0.50, on the odds-ratio scale for different precisions around the variance and different anticipated responses for 90% power and type I error of 2.5%

Control Rate	Odds-Ratio	Degrees of Freedom					
		10	20	30	40	50	100
0.80	0.70	1344	1177	1130	1108	1095	1069
	0.80	726	631	605	593	585	571
	0.90	526	442	419	408	402	389
	1.00	494	406	382	371	364	351
	1.10	558	463	437	425	418	404
	1.20	723	615	585	572	563	548
	1.30	1028	881	840	821	810	787
	1.40	1536	1314	1253	1224	1207	1174
0.85	0.70	1945	1541	1445	1402	1377	1329
	0.80	1069	836	782	757	743	716
	0.90	807	600	552	530	518	494
	1.00	774	560	510	487	474	449
	1.10	867	637	583	559	545	518
	1.20	1101	836	775	748	732	702
	1.30	1551	1196	1113	1075	1054	1012
	1.40	2322	1791	1665	1608	1575	1513
0.90	0.70	4584	2487	2193	2073	2008	1888
	0.80	2635	1368	1200	1132	1095	1027
	0.90	2159	1020	872	812	779	718
	1.00	2132	973	821	758	724	660
	1.10	2350	1097	933	866	830	762
	1.20	2850	1411	1223	1147	1105	1030
	1.30	3838	2005	1751	1647	1590	1487
	1.40	5682	3006	2625	2469	2384	2228

4.5.1.20. Worked Example

Suppose the control response rate was estimated from 100 patients. Repeating the same calculations from earlier for the same equivalence limit of OR=0.5 the sample size should be increased to 449 patients per arm around a 6% increase in the sample size.

Repeating the sample calculations on the absolute difference scale increases the same size to 156 patients per arm. This is an increase in the sample size of 5%.

Table 4-34. Sample sizes for an equivalence study on the absolute difference scale for different precisions around the variance and different anticipated responses for 90% power and type I error of 2.5%

Control Rate	p_A-p_B	Limit	Degrees of Freedom 10	20	30	40	50	100
0.80	0.10	0.15	1564	1569	1564	1564	1564	1560
	-0.05	0.10	1495	1487	1480	1480	1474	1469
		0.15	374	374	370	370	370	367
	0.00	0.05	1852	1782	1753	1728	1717	1698
		0.10	469	450	442	435	429	424
		0.15	211	198	198	193	193	189
	+0.05	0.10	1327	1273	1262	1243	1243	1228
		0.15	337	322	316	311	311	307
	+0.10	0.15	1220	1154	1127	1104	1094	1077
0.85	-0.10	0.15	1400	1377	1367	1358	1350	1337
	-0.05	0.10	1327	1285	1262	1252	1243	1228
		0.15	337	322	316	316	311	307
	0.00	0.05	1649	1534	1477	144	1415	1379
		0.10	421	385	366	358	358	345
		0.15	182	174	167	161	161	156
	+0.05	0.10	1144	1049	1004	991	968	949
		0.15	293	263	255	248	242	237
0.90	-0.05	0.10	1123	1033	1004	979	968	941
	0.00	0.05	1399	1234	1144	1105	1070	1012
		0.10	352	310	287	277	268	253

4.5.1.21. Calculations that take Account of the Imprecision in the Estimates of the Effects Used in the Sample Size Calculations – Bayesian Methods

As discussed previously in this chapter the percentiles for a posterior control response can be calculated as described in 4.1.1.16 to give an estimate of the sample size. As previously it is best to highlight the points through worked example.

4.5.1.22. Worked Example

For the absolute difference scale with a non-informative prior the sample size is estimate to be 153 patients per arm. This is 3 less than the sample size calculated before.

With a more pessimistic prior (the most likely response being 80% with 90% certainty that it is greater than 75%), the sample size estimate is increased to 173 patients per arm.

With a prior that the control response rate observed is about right (the most likely response being 85% with 90% certainty it is greater than 80%) the sample size estimate is increased to 152 patients per arm.

Similar calculations could be done if equivalence is defined in terms of an odds-ratio

4.5.1.23. Calculations Taking Account of the Imprecision of the Populations Effects With Respect to the Assumptions about the Mean Difference and the Variance Used in the Sample Size Calculations

To allow for the imprecision in the assumptions about both the mean difference and variance numerical methods could be used to calculate the sample size on the absolute difference scale from.

$$1-\beta = \frac{1}{0.998}\sum_{perc=0.001}^{0.998}\frac{\lambda_1+\lambda_2}{2}, \qquad (4.3.20)$$

where λ_A and λ_B are defined as

$$\lambda_1 = \Phi\left(\sqrt{\frac{n\left(\left(p_{perc_A}-p_B\right)-d\right)^2}{\left(p_{perc_A}(1-p_{perc_A})+p_B(1-p_B)\right)}}-Z_{1-\alpha}\right)+\Phi\left(\sqrt{\frac{n\left(\left(p_{perc_A}-p_B\right)-d\right)^2}{\left(p_{perc_A}(1-p_{perc_A})+p_B(1-p_B)\right)}}-Z_{1-\alpha}\right)-1,$$

$$\lambda_2 = \Phi\left(\sqrt{\frac{n\left(\left(p_{perc_A}-p_B\right)-d\right)^2}{\left(p_{(perc+0.001)_A}(1-p_{(perc+0.001)_A})+p_B(1-p_B)\right)}}-Z_{1-\alpha}\right)+\Phi\left(\sqrt{\frac{n\left(\left(p_{perc_A}-p_B\right)-d\right)^2}{\left(p_{(perc+0.001)_A}(1-p_{(perc+0.001)_A})+p_B(1-p_B)\right)}}-Z_{1-\alpha}\right)-1.$$

Note that, similar to non-inferiority calculations given earlier, a number issues need to be additionally considered

1. If any of the percentile values around the control response crosses a step (given in Table 4.23) then the non-inferiority margin, d, should be altered accordingly.

2. The investigative response rate, p_B, remains assumed fixed, calculated from the initial p_A, and not from individual p_{perc_A}.

3. Following on from 2. for instances where $p_{perc_A} - p_B$ exceeds an equivalence bound then the power for this percentile (to be averaged across for power calculation) is set to 0.

The equivalent calculation to estimate the sample size for an equivalence study design based around the odds-ratio would be

$$1-\beta = \frac{1}{0.998}\sum_{perc=0.001}^{0.998}\frac{\eta_1+\eta_2}{2}, \qquad (4.3.21)$$

where η_A and η_B are defined as

$$\eta_1 = \Phi\left(\sqrt{n(\log(OR_{perc})-d)^2\left[1-\sum_{i=1}^{2}\bar{p}^3_{perc}\right]/6}-Z_{1-\alpha}\right)+\Phi\left(\sqrt{n(\log(OR_{perc})+d)^2\left[1-\sum_{i=1}^{2}\bar{p}^3_{perc_i}\right]/6}-Z_{1-\alpha}\right)-1$$

$$\eta_2 = \Phi\left(\sqrt{n(\log(OR_{perc})-d)^2\left[1-\sum_{i=1}^{2}\bar{p}^3_{(perc+0.001)}\right]/6}-Z_{1-\alpha}\right)+\Phi\left(\sqrt{n(\log(OR_{perc})+d)^2\left[1-\sum_{i=1}^{2}\bar{p}^3_{(perc+0.001)_i}\right]/6}-Z_{1-\alpha}\right)$$

As the odds-ratio does not suffer from the issues of stepped equivalence bounds, the calculations are more straightforward. The following two points should be considered however

1. The investigative response rate, p_B, remains fixed and is estimated from the initial p_A.

2. Following on from 2. for instances where $OR_{perc} = (p_{perc}(1-p_B))/(p_B(1-p_{perc}))$ exceeds an equivalence bound then the power for this percentile (to be averaged across for power calculation) is set to 0.

Tables 4.35 and 4.36 give sample size calculations using (4.3.21) and (4.3.20) respectively. As with non-inferiority trials discussed earlier in this chapter the calculations for the odds-ratio give, what could be considered to be, unfeasibibly large sample size estimates. Following on from the same arguments for non-inferiority trials the recommendation therefore is to use (4.3.18) and Table 4.33 when working on the odds-ratio scale and (4.3.20) and Table 4.36 when working with absolute differences.

Table 4-35. Sample sizes for an equivalence study, with a limit of 0.5, on the odds-ratio scale for different precisions around the anticipated control response rate and variance for 90% power and type I error of 2.5%. The true odds-ratio is fixed at 1.00

Control	Degrees of Freedom			
Rate	50	100	250	500
0.70	1031	464	321	287
0.75	1486	563	369	325
0.80	2833	758	452	389
0.85	12510	1242	614	507
0.90	NE	3524	1034	776

NE – Not estimable

4.5.1.24. Worked Example

Repeating the worked example from earlier where the control response rate was estimated from 100 patients the sample calculations on the absolute difference scale increases the sample size to 194 patients per arm.

Table 4-36. Sample sizes for an equivalence study on the absolute difference scale for different precisions around the anticipated control response rate and variance for 90% power and type I error of 2.5%

Control Rate	p_A-p_B	Limit	Degrees of Freedom 50	100	250	500
0.80	-0.10	0.15	11554	11554	4075	2400
	-0.05	0.10	1460	1460	1460	1460
		0.15	1014	583	417	365
	0.00	0.05	1663	1663	1663	1663
		0.10	415	415	415	415
		0.15	269	199	184	184
	+0.05	0.10	1208	1208	1208	1208
		0.15	349	302	302	302
	+0.10	0.15	1050	1050	1050	1050
0.85	-0.10	0.15	11322	1129	2640	1842
	-0.05	0.10	11207	1208	1208	1208
		0.15	10301	465	345	323
	0.00	0.05	1325	1325	1325	1325
		0.10	356	331	331	331
		0.15	356	194	165	157
	+0.05	0.10	914	914	914	914
		0.15	418	330	290	262
0.90	-0.05	0.10	10913	2797	1363	1067
	0.00	0.05	935	935	935	935
		0.10	324	246	233	233

4.5.1.25. Calculations Taking that take Account of the Imprecision of the Populations Effects With Respect to the Assumptions about the Mean Difference and the Variance Used in the Sample Size Calculations – Bayesian Methods

In the following worked example the calculations are repeated using Bayesian methods to estimate posterior percentiles to use in (4.3.21).

4.5.1.26. Worked Example

For the absolute difference scale with a non-informative prior the sample size is estimated to be 201 patients per arm. This is 7 greater than the sample size calculated before.

With a more pessimistic prior (the most likely response being 90% with 90% certainty that it is greater than 85%), the sample size estimate is increased to 233 patients per arm.

With a prior that the control response rate observed is correct (the most likely response being 85% with 90% certainty it is greater than 80%) the sample size estimate is 168 patients per arm.

4.5.2. Cross-over Trials

As with non-inferiority trials discussed earlier in this chapter the arguments for superiority and non-inferiority trials can be extended to equivalence trials. Although there are a number of paper which deal specifically on this topic [Tango 1998, 1999; Nam, 1977; Tang, Tang and Chan, 2003], it is recommended to use the parallel group sample size methodologies per arm in this sub section of the chapter to estimate the total sample size for an equivalence cross-over trial

4.6. Estimation to a Given Precision

4.6.1. Parallel Group Trials

4.6.1.1. Sample Sizes with the Population Effects Assumed Known

4.6.1.2. Proportional Difference

In a two-group study where the primary outcome is binary and the objective is to estimate the possible population difference such that

$$d = p_A - p_B,$$

where p_A and p_B are sample proportional responses on treatment groups A and B respectively. As discussed in Chapters 1 and 3 a $(1-\alpha)100\%$ Normal approximation confidence interval for f(μ) has half-width

$$w = Z_{\alpha/2}\sqrt{Var(S)}, \qquad (4.4.1)$$

where var(S) is defined as

$$Var(S) = \frac{p_A(1-p_A) + p_B(1-p_B)}{n}, \qquad (4.4.2)$$

which can in turn be approximated from

$$Var(S) \approx \frac{2\bar{p}(1-\bar{p})}{n}, \qquad (4.4.3)$$

where $\bar{p} = (p_A + p_B)/2$ i.e. the mean proportion expected across both the treatments. Remember from earlier in the chapter that from Table 4.19 it seems that this approximate variance formula holds for proportional responses (p_A and p_B) that are within +/-0.30 of each other and thus covers most practical situations. For trials based on precision considerations, therefore, it may be optimal to use an estimate of the mean overall response for the variance and subsequent sample size calculations (given that it is probably the aim of the study to estimate possible treatment responses). However, if one has reasonable estimates for each treatment response then these should be used in calculations.

Therefore, for a given half confidence interval width, w the following condition must be met to obtain the sample size per group

$$n = \frac{2\bar{p}(1-\bar{p})Z^2_{1-\alpha/2}}{w^2}. \qquad (4.4.4)$$

From (4.4.4) Table 4.37 is derived. Table 4.37 gives the sample size required for different values of the expected mean response across treatment groups, $\bar{p}$, and widths w. Two sided 95% confidence intervals are assumed to be planned to be constructed. The mean responses, $\bar{p}$, given in the table vary from 0.10 to 0.50. Values greater than 0.50 are not given as the sample size required for $\bar{p}$=0.60 is equivalent to $\bar{p}$=0.40, the sample size for $\bar{p}$=0.70 is the same as $\bar{p}$=0.30 etc.

Table 4-37. Sample sizes required per group for two sided 95% confidence intervals for different values of width, w, for various expected mean proportional responses

	w				
$\bar{p}$	5	10	15	20	25
0.10	277	70	31	18	12
0.20	492	123	55	31	20
0.30	646	162	72	41	26
0.40	738	185	82	47	30
0.50	769	193	84	49	31

4.6.1.3. Odds-Ratio

For binary data the difference in the sample proportions may also be expressed terms of an odds-ratio (OR)

$$d = OR = \frac{p_A(1-p_B)}{p_B(1-p_A)}.$$

A $(1-\alpha)100\%$ confidence interval for log(d) can be derived using the following variance estimate [Whitehead, 1993]

$$Var(\log(d)) = \frac{6}{n\left[1-\sum_{i=1}^{2}\bar{p}_i^3\right]}.$$

Therefore, as for binary data for a given half confidence interval width, w, around the odds-ratio the following condition must be met to obtain the sample size per group

$$n = \frac{6\,Z^2_{1-\alpha/2}}{(\log(1-w))^2\left[1-\sum_{i=1}^{2}\bar{p}_i^3\right]}, \qquad (4.4.5)$$

where $\bar{p}_i$ are the expected mean responses. Remember that for binary data $\bar{p}_1 = (p_A + p_B)/2 = \bar{p}$, say, and $\bar{p}_2 = 1 - \bar{p}_1 = 1 - \bar{p}$ and thus correspond to $\bar{p}$ given earlier in this chapter.

Note that in this instance w is on the log scale and so w=0.60 would equate to a confidence interval for a given odds-ratio, OR, being within (1-w)OR to OR/(1-w) i.e. 0.40OR to 2.5OR. Also note that log(1-w) is on the arithmetic scale i.e. log(1-w)=-log[1/(1-w)]

Table 4.38 gives sample size required for different values of the mean response across treatment groups, $\bar{p}$, and widths w estimated using (4.4.5). Two sided 95% confidence intervals are again assumed. As with Table 4.36 the mean responses, $\bar{p}$, given in the table vary from 0.10 to 0.50. To obtain a sample size for $\bar{p}$ >0.5 look up 1- $\bar{p}$.

Table 4-38. Sample sizes required per group for two sided 95% confidence intervals for different values of width w around the odds-ratio for various expected mean proportional responses

	w					
$\bar{p}$	0.25	0.30	0.35	0.40	0.45	0.50
0.10	1032	672	461	328	239	178
0.20	581	378	259	185	135	100
0.30	443	288	198	141	103	77
0.40	387	252	173	123	90	67
0.50	372	242	166	118	86	64

A feature of Table 4.38 that is different to Table 4.37 but which is consistent with the other types of trial described in this chapter, is that as $\bar{p}$ approaches 0.5 the smaller sample size that is required for equivalent values of w. Table 4.39 further illustrates this. For different mean response rates and values of w around the odds ratio the equivalent widths on the proportional scale are given. For example for a w=0.50 on the odds-ratio scale an equivalent width would be 0.173 on the proportional scale for $\bar{p}$=0.50. Details of the derivation of Table 4.39 are given in the next sub-section.

4.6.1.4. Equating Odds-Ratios with Proportions

As with superiority trials discussed earlier (4.4.4) and (4.4.5) can be approximately equated. If one redefined the half widths around the confidence intervals of the odds-ratios (w_{or}) and proportional differences (w_p) in terms of odds-ratios and proportional differences one obtains respectively

$w_p = (p_A - p_B) - (p_{A_l} - p_{B_l})$,

$$1 - w_{or} = OR/OR_L = \frac{p_A(1-p_B)}{p_B(1-p_A)} \Big/ \frac{p_{A_l}(1-p_{B_l})}{p_{B_l}(1-p_{A_l})}.$$

Thus, (4.4.4) and (4.4.5) can be re-written as

$$n = \frac{2\bar{p}(1-\bar{p})Z^2_{1-\alpha/2}}{\left[(p_A - p_B) - (p_{A_l} - p_{B_l})\right]^2}, \tag{4.4.6}$$

$$n \geq \frac{6\,Z^2_{1-\alpha/2}/(\log OR - \log \mathrm{OR}_L)^2}{\left[1 - \sum_{i=1}^{2} \bar{p}_i^3\right]}. \tag{4.4.7}$$

Remembering that,

$$\frac{6}{\left(1 - \sum_{i=1}^{2} \bar{p}_i^3\right)} = \frac{2}{\bar{p}(1-\bar{p})}, \tag{4.4.8}$$

and $\log(OR) \approx 2(OR-1)/(OR+1)$, which holds for odds-ratios within $0.33 \leq \mathrm{OR} \leq 3.00$, hence,

$$\frac{2(OR-1)}{OR+1} \approx \frac{p_A - p_B}{\bar{p}(1-\bar{p})}, \tag{4.4.9}$$

and

$$\log OR_L \approx \frac{p_{A_l} - p_{B_l}}{\bar{p}(1-\bar{p})} \tag{4.4.10}$$

Assuming $\bar{p}(1-\bar{p}) \approx \bar{p}_L(1-\bar{p}_L)$ where $\bar{p}_L = (p_{A_l} + p_{B_l})/2$ and substituting (4.4.8), (4.4.9) and (4.4.10) into (4.4.5) one obtains

$$n \geq Z^2_{1-\alpha/2} \frac{2}{\bar{p}(1-\bar{p})}\left(\frac{\bar{p}(1-\bar{p})}{(p_A - p_B) - (p_{A_l} - p_{B_l})}\right)^2 = \frac{2\bar{p}(1-\bar{p})Z^2_{1-\alpha/2}}{\left[(p_A - p_B) - (p_{A_l} - p_{B_l})\right]^2}.$$

Thus, similarly to superiority trials, (4.4.4) and (4.4.5) can be used interchangeably depending on preference. Due to this property one therefore has

$$\frac{2\bar{p}(1-\bar{p})Z^2_{1-\alpha/2}}{(w_p)^2} \approx \frac{2Z^2_{1-\alpha/2}}{(\log(1-w_{OR}))^2\,\bar{p}(1-\bar{p})},$$

and thus

$$w_p \approx \left|\log(1-w_{OR})\right|\left(\bar{p}(1-\bar{p})\right). \tag{4.4.11}$$

Using (4.4.11) Table 4.39 can be derived.

Table 4-39. Table of widths on the absolute difference scale that are equivalence to the widths, w, around the odds-ratio for various anticipated expected mean Proportions

	w					
$\bar{p}$	0.25	0.30	0.35	0.40	0.45	0.50
0.10	0.026	0.032	0.039	0.046	0.054	0.062
0.20	0.046	0.057	0.069	0.082	0.096	0.111
0.30	0.060	0.075	0.090	0.107	0.126	0.146
0.40	0.069	0.086	0.103	0.123	0.143	0.166
0.50	0.072	0.089	0.108	0.128	0.149	0.173

4.6.1.5. Worked Example

A pilot study is planned to estimate the odds ratio between comparator and control regimens. The expected mean response rate is 50% across the two treatments and the wish is to quantify the odds ratio within ±55% (i.e. w=55%). This means that if a odds ratio of 0.70 was observed, one would be able to say that the true odds-ratio is likely to be between 0.32 and 1.56. Therefore, from Table 4.38 the sample size required is 49 per group.

Following from the example from Table 4.39 w=0.55 on the odds ratio scale is equivalent to proportional half confidence width of 20% for a mean response rate of 50%. From Table 4.38 a width of 20% and a mean proportional response of 50% gives 49 subjects per group again.

4.6.1.6. Sensitivity Analysis About the Estimates of the Population Effects Used in the Sample Size Calculations

Extending the arguments for other types trial discussed earlier in this chapter the sensitivity of a precision based can be investigated through construction of a 95% confidence interval around the anticipated overall response rate. For each tail of this confidence interval one can re-investigate the precision of the trial to give a quantification of its sensitivity.

4.6.1.7. Worked Example

Suppose the expected response rate of 50% for the worked example earlier was estimated from a trial with 50 patients. The 95% confidence interval for this would be between 33.2% and 66.8%.

On the absolute difference scale these lower and upper tails would give precision of 19%. A slight improvement over the previous calculations – due to 50% giving the maximum variance estimate.

On the odds ratio scale the precision for each tail would be 57%, which is a little worse than that previously observed.

4.6.1.8. Calculations Taking of Account the Imprecision of the Estimates of the Population Effects Used in the Sample Size Calculations

Adapting the formula for superiority trials the following result could be used to calculate sample sizes to account for the imprecision in the sample variance on the odds-ratio scale

$$\frac{1}{0.998}\sum_{p_{pre}=0.001}^{0.999} 0.5\left[\begin{array}{l}\Phi\left(\sqrt{n(\log(1-w))^2\left[1-\sum_{j=1}^{2}p_{pre}^3\right]/6}-Z_{1-\alpha/2}\right)\\ +\Phi\left(\sqrt{n(\log(1-w))^2\left[1-\sum_{j=1}^{2}p_{(1/or)pre}^3\right]/6}-Z_{1-\alpha/2}\right)\end{array}\right]\geq 0.5. \qquad (4.4.12)$$

Table 4.40 gives the sample sizes using (4.4.12) for difference precisions around the sample variance. For trials based on estimated it seems that the imprecision around the sample variance has little impact on the sample size estimate.

Table 4-40. Sample sizes for a precision based study, for precision of width, w, of 0.50, on the odds-ratio scale for different precisions around the variance and different anticipated overall responses for a 95% confidence interval

Overall	Degrees of Freedom					
Rate	10	20	30	40	50	100
0.10	186	179	177	177	177	177
0.20	108	104	104	104	101	101
0.30	84	80	80	80	80	77
0.40	75	71	71	71	68	68
0.50	72	68	68	68	68	65

Nominally the equivalent result if working on the absolute difference scale would be

$$\frac{1}{0.998}\sum_{p_{pre}=0.001}^{0.999} 0.5\left[\begin{array}{l}\Phi\left(\sqrt{\frac{n(p_1-p_R)^2}{p_{pre}(1-p_{pre})+p_{(1/or)pre}(1-p_{(1/or)pre})}}-Z_{1-\alpha/2}\right)+\\ \Phi\left(\sqrt{\frac{n(p_1-p_R)^2}{p_{pre}(1-p_{pre})+p_{(1/or)pre}(1-p_{(1/or)pre})}}-Z_{1-\alpha/2}\right)\end{array}\right]\geq 0.5, \qquad (4.4.13)$$

however, accounting for the imprecision in the variance has no effect on the sample size when working on this scale and so no table is given and this result can be ignored.

4.6.1.9. Worked Example

Suppose an investigator wished to design a trial where the anticipated overall response rate is 50% and the investigator wished to quantify the odds ratio with precision of 0.5. From Table 4.40 the sample size required is 64 patients per group.

Now supposing the overall response was estimated from a trial with 30 patients. Accounting for this would increase the sample size to 68 patients.

4.6.1.10. Calculations that take Account of the Imprecision in the Estimates Used in the Sample Size Calculations – Bayesian Methods

As discussed earlier in this chapter the percentiles for a posterior control response can be calculated as described in 4.1.1.16 to give an estimate of the sample size. As previously it is best to highlight the points through a worked example.

4.6.1.11. Worked Example

If initially a non-informative prior was used then the sample size is estimated at 67 patients per arm. A sample size 1 short of calculated previously.

Supposing the investigator was sceptical as the control response being as high as 50% such that the belief was that the most likely response was 40% with at least 90% certainty that it was greater than 30%. The estimate of sample size is now 66.

4.6.2. Cross-Over Trials

As with the other types of trial discussed it is recommended that the total sample size for a cross-over precision based trial be taken from the one arm sample size for a parallel group trial. There are a number papers that discuss issues with confidence intervals for cross-over trials [Tango, 1998, 1999; Newcombe, 1998c; May and Johnson, 1997].

4.7. Design Considerations

4.7.1. Inclusion of Baselines or Covariates

As discussed in chapter 2 in the analysis of the results of a clinical trial, the effects of treatment on the response of interest are often adjusted for predictive factors, such as demographic or clinical covariates by fitting them concurrently with the treatment variable. It was highlighted in this chapter how when adjusting for a highly predictive covariate, such as baseline, the sample size can be dramatically reduced due a reduction in the variance estimate.

It has previously been reported that for the case of binary data when adjusting for a baseline or other predictive covariate the variance estimate increases [Whitehead, 1993; Robinson and Jewell, 1991]. The inference from this is that for binary data variance adjustment increases the sample size and not decreases it as for Normal data. This point will now be challenged.

Suppose one is designing a trial to compare two treatments where a binary baseline covariate has been assessed. The relationship between the covariate and outcome is assumed to be an odds-ratio, OR_C. While the relationship between the treatment, and the outcome, the treatment effect, is assumed to be OR_T.

Assuming that final analysis is to be through a logistic regression the model would be defined as

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{Constant} + \text{B} \times \text{covariate(0 or 1)} + \text{C} \times \text{treatment(0 or 1)}, \qquad (4.5.1)$$

or

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta .$$

Hence,

$$p_i = \frac{1}{1+e^{-\beta}} \quad . \qquad (4.5.2)$$

Now, for the case of the covariate taking level 0 the probability of observing outcome level 1 for each treatment (A and B say for levels 0 and 1 respectively) can be defined as

$$p_{A1} = \frac{1}{1+A} \text{ and } p_{B1} = \frac{1}{1+OR_T A} = \frac{OR_T^{-1}}{A+OR_T^{-1}} . \qquad (4.5.3)$$

Note $A = e^{\text{Constant}}$. Remember, that the variance for the log-odds-ratio is defined as

$$\text{var}(\log(OR)) = \frac{6}{n\left(1-\sum_{i=1}^{2}\bar{p}_i^3\right)} = \frac{2}{n\bar{p}_1(1-\bar{p}_1)}, \qquad (4.5.4)$$

where $\bar{p}_i$ is defined as the average response across treatments for each outcome category ($\bar{p}_1 = (p_{A1} + p_{B1})/2$ and $\bar{p}_2 = 1 - \bar{p}_1$). Hence,

$$\bar{p}_1 = \frac{2 + A(OR_T + 1)}{2(1+A)(1+OR_T A)}, \qquad (4.5.5)$$

and

$$\bar{p}_1(1-\bar{p}_1)=\left(\frac{((1+2A)OR_I+1)A}{2(1+A)(1+OR_IA)}\right)\left(\frac{2+A(OR_T+1)}{2(A+1)(1+OR_TA)}\right). \qquad (4.5.6)$$

Equivalently, for the case where the covariate takes level 1 the probability of observing outcome level 1 for each treatment can be defined as

$$p_{A1}=\frac{OR_C}{A+OR_C} \text{ and } p_{B1}=\frac{OR_COR_I}{A+OR_COR_I}. \qquad (4.5.7)$$

Hence,

$$\bar{p}_1=\frac{OR_C(A+OR_COR_I)+(A+OR_C)OR_COR_I}{2(OR_C+A)(A+OR_COR_I)}, \qquad (4.5.8)$$

and

$$\bar{p}(1-\bar{p})=\left(\frac{OR_C(A+OR_COR_I)+(A+OR_C)OR_COR_T}{2(OR_C+A)(A+OR_COR_I)}\right)\left(\frac{A(2A+OR_COR_T+OR_C)}{2(A+OR_C)(A+OR_T)}\right). \qquad (4.5.9)$$

To obtain an overall estimate of effect across the two levels of the covariate one could use

$$\log(OR)=\frac{\sum_{i=0}^{k}w_i\log(OR_i)}{\sum_{i=0}^{k}w_i}, \qquad (4.5.10)$$

where OR_i is an estimate of the response from covariate level 1 and w_i is the reciprocal of the variance from covariate level i ($w_i=1/\text{var}(\log(OR_i))$) and k is the number of levels for the covariate. Consequently, an overall estimate of the variance is defined as

$$\frac{1}{\sum_{i=0}^{k}w_i}. \qquad (4.5.11).$$

Hence, for the case of a two level covariate described here one has

$$w_0=\frac{8(A+1)^2(A+OR_I)^2}{((A+OR_I)+(A+1)OR_I)A(2A+OR_I+1)},$$

$$w_1=\frac{(OR_CA((2OR_CA+1)OR_I+1))(2+OR_CA(1+OR_T))}{8(OR_CA+1)^2(1+OR_COR_IA)^2}. \qquad (4.5.12)$$

This can be put into (4.5.11) to obtain an overall estimate of the variance. Note that for the special case of $OR_C=1$ one has $w_0=w_1$. Note that in all of these calculations the assumption is that there is no interaction between treatment and the covariate i.e. the effect of treatment is independent of covariate.

Table 4-41. Bias and variance inflation for unadjusted logistic regression for various odd-ratios for the covariate and treatment

Covariate OR	Treatment OR	Pooled OR*	SE Pooled OR	Unadjust OR	Unadjust SE OR	SE ratio Unadj:Pool	Bias Unad:Pool	Pooled Standard	Unadjust Standard	Stan Ratio Unadj:Pool
0.60	0.50	0.50	0.2098	0.5051	0.2084	0.9929	0.9854	3.3032	3.2779	0.9923
	0.60	0.60	0.2077	0.6046	0.2062	0.9926	0.9849	2.4595	2.4405	0.9923
	0.70	0.70	0.2061	0.7038	0.2045	0.9924	0.9846	1.7308	1.7173	0.9922
	0.80	0.80	0.2048	0.8028	0.2033	0.9922	0.9844	1.0893	1.0808	0.9921
	0.90	0.90	0.2039	0.9015	0.2023	0.9921	0.9842	0.5167	0.5126	0.9921
0.70	0.50	0.50	0.2072	0.5025	0.2064	0.9965	0.9927	3.3458	3.3330	0.9962
	0.60	0.60	0.2053	0.6023	0.2045	0.9963	0.9925	2.4887	2.4791	0.9961
	0.70	0.70	0.2039	0.7019	0.2031	0.9962	0.9924	1.7496	1.7428	0.9961
	0.80	0.80	0.2028	0.8014	0.2021	0.9961	0.9923	1.1001	1.0958	0.9961
	0.90	0.90	0.2021	0.9007	0.2013	0.9961	0.9922	0.5213	0.5193	0.9961
0.80	0.50	0.50	0.2052	0.5010	0.2049	0.9986	0.9971	3.3775	3.3723	0.9985
	0.60	0.60	0.2035	0.6009	0.2033	0.9985	0.9970	2.5096	2.5058	0.9985
	0.70	0.70	0.2024	0.7008	0.2021	0.9985	0.9970	1.7626	1.7598	0.9985
	0.80	0.80	0.2015	0.8005	0.2012	0.9985	0.9969	1.1072	1.1055	0.9985
	0.90	0.90	0.2010	0.9003	0.2007	0.9985	0.9969	0.5242	0.5234	0.9985
0.90	0.50	0.50	0.2038	0.5002	0.2038	0.9997	0.9993	3.4006	3.3995	0.9997
	0.60	0.60	0.2024	0.6002	0.2023	0.9997	0.9993	2.5243	2.5234	0.9997
	0.70	0.70	0.2014	0.7002	0.2013	0.9997	0.9993	1.7711	1.7705	0.9997
	0.80	0.80	0.2007	0.8001	0.2007	0.9997	0.9993	1.1116	1.1112	0.9997
	0.90	0.90	0.2003	0.9001	0.2003	0.9997	0.9993	0.5259	0.5257	0.9997
1.00	0.50	0.50	0.2028	0.5000	0.2028	1.0000	1.0000	3.4173	3.4173	1.0000
	0.60	0.60	0.2016	0.6000	0.2016	1.0000	1.0000	2.5341	2.5341	1.0000
	0.70	0.70	0.2008	0.7000	0.2008	1.0000	1.0000	1.7764	1.7764	1.0000
	0.80	0.80	0.2003	0.8000	0.2003	1.0000	1.0000	1.1140	1.1140	1.0000
	0.90	0.90	0.2001	0.9000	0.2001	1.0000	1.0000	0.5266	0.5266	1.0000

* - The pooled OR is the odds-ratio estimated through combining the ORs from each sub-group (which are equal)

Unadjusted estimates of the overall effects can be obtained from averaging the estimates of p_{A1} and p_{A1} from (4.5.3) and (4.5.7) – note here one is assuming equal sample size for each level of the covariate. Hence, one has

$$p_{A1} = \frac{1}{2}\left(\frac{1}{1+A} + \frac{1}{OR_C A + 1}\right) = \frac{(1+OR_C A)+(A+1)}{2(1+A)(1+OR_C A)}.$$

$$p_{B1} = \frac{1}{2}\left(\frac{1}{1+OR_I A} + \frac{1}{1+OR_C OR_I A}\right) = \frac{(1+OR_C OR_I A)+(1+OR_T A)}{2(1+OR_I A)(1+OR_C OR_T A)}. \qquad (4.5.13)$$

Obviously, an unadjusted estimated of the log-odds-ratio can be estimated from (4.5.13) from

$$\log(OR) = \frac{p_{A1}(1-p_{B1})}{p_{B1}(1-p_{A1})}. \qquad (4.5.14)$$

The variance for which can be estimated through putting (4.5.13) into (4.5.4)

Using (4.5.14) for unadjusted estimates of the overall response and (4.5.13) with (4.5.4) for unadjusted estimates of the variance and (4.5.10) and (4.5.11) for adjusted estimates of the overall response and variance respectively Table 4.41 was constructed (note A was fixed at 1) with the sample size fixed at 100 per group. The first column gives the OR estimate for the covariate effect with respect to outcome and the second column gives the OR estimate for the treatment effect with respect to outcome.

Column three gives the adjusted point estimate from (4.5.10) and column four its corresponding standard error. Column 5 and 6 give the corresponding unadjusted point estimates with their standard errors. It seems from inspection of these results that the variance estimate is indeed bigger when adjusting and is confirmed in column 7, which gives the ratio of the standard errors. However, this result should be put into context with the next column, which gives a ratio of the log-odd-ratios. From this column it seems that although the unadjusted estimates have smaller errors the point estimates are biased and this bias is towards the "null" hypothesis. The bias in not adjusting for covariates has previously been commented on [Gail, Wieand and Paintadose, 1984] and from these results it seems that the bias gets bigger the bigger the effect of the covariate.

Columns 9 and 10 further confirm these points as these give the Z-statistic of the point estimate (log-odds-ratio) divided by it standard error. From this one can see that the Z-statistic is consistently smaller for the unadjusted estimates compared to the adjusted (as evidenced by the final column which gives the ratio of the of the Z-statistics).

From the results of Table 4.41 it is evident that relative to the point estimate (the Z-statistic) adjusting for a covariate does not increase the variance. In fact through adjusting for covariates one is obtaining unbiased estimates of the treatment effect. Practically what

these results highlight is that adjusting for covariates does not (relatively) increase variance estimate and so one does not need to increase the sample size to account for any planned covariate adjustment.

Note in the hypothetical example in Table 4.18 the unadjusted log-odds-ratio has a smaller standard error of 0.119 compared to 0.125 for a logistic regression analysis in PROC LOGISTIC in SAS adjusting for gender. Hence, the standard error has increased by 5% through covariate adjustment. However, this effect is more than offset by the bias in the log odds ratio, with the unadjusted log-odds-ratio being 1.022 (odds-ratio=2.78) compared to an adjusted 1.099 (odds-ratio=3.00). A 7.5% increase in the log-odds-ratio by adjusting. Using (4.5.1) to estimate the standard error for Table 4.18 gives an estimate of 0.119 for the adjusted analysis and (4.5.4) gives an estimate of 0.116 both a little lower than using PROC LOGISTIC. The estimates from (4.5.10) and (4.10.14) match exactly the point estimates from PROC LOGISTIC

As an aside it may be worth noting what ICH E9 [1998] says for the case where one is unsure of the effect of a covariate *a priori*

"When the potential value of an adjustment is in doubt, it is often advisable to nominate the unadjusted analysis as the one for primary attention, the adjusted being supportive"

Given the direction of the bias this advice may be true for superiority trials with binary data as the bias is towards the "null" and would be against the investigative treatment. However, for non-inferiority and equivalence trials the aim is to demonstrate no effect and so an unadjusted analysis would be biased in favour of the investigative treatment. The biased estimate would also have a smaller standard error too which would be most optimal for demonstrating a confidence interval is contained within a bound. The advice in E9 should therefore take the opposite standpoint for binary non-inferiority and equivalence trials i.e. the adjusted should be the primary analysis.

The situation of what is conservative for superiority trials being counter conservative for equivalence or non-inferiority trials is recognised in other aspects of study conduct – for example whether the pre protocol or intent to treat population is the primary analysis set – but to date no recognition with respect to covariates and logistic regression.

4.7.2. Post Dose Measures Summarised by Summary Statistics

There have been a number of articles written on the topic of repeated binary data [Nixon and Thompson, 2003; Liu, 1991, 2001], however, here all that will be highlighted is how, for the special case where the post dose measures are analysed using summary statistics, the form of the data changes.

For the simple case of two post dose measurements, by taking a simple average of the two measures the scale moves from a two-point scale to a three-point scale. The data has thus been transformed from binary to ordinal in form. The advantage with reference to the sample size can be observed in Table 4.42. From this table it is evident that if there are several post dose measurements the sample size can be reduced by up to a third (compared to a simple binary response).

Discussion of Table 4.42 will be made in greater detail in Chapter 5 in context with an ordinal response.

Table 4-42. Correction factors to use when the number of categories is less than or equal to 5

Number of Categories	Correction Factor
2	1.333
3	1.125
4	1.067
5	1.042

4.8. Summary of Chapter 4

For cross-over trials it is recommended that the standard parallel group methodologies described in this chapter be used to calculate the sample size, taking the sample size per group from these calculations as the total sample size. It is recommended that the effect of period be accounted for in any statistical analysis but that there is no need for it to be considered when calculating a sample size.

For parallel group trials it was highlighted how not allowing for a predictive covariate in a statistical analysis could bias one's results and it was demonstrated how when adjusting for these covariates that it did not inflate one's sample size. It was highlighted that an unadjusted analysis was a conservative analysis for a superiority trial but not for a non-inferiority or equivalence trial.

When designing a study with a binary response one should account for the imprecision around the estimated control response rate. This recommendation is particularly important for either high or low anticipated response rates. It is recommended numerical methods derived in this chapter be used to account for the imprecision in the control response over the results described for Normal data in Chapter 2.

For binary data simple Bayesian procedures can considerably enhance the estimates of the sample size. As well as using empirical observation, beliefs about the results that are

anticipated can be used in sample size calculations. The use of a priors provides an opportunity to the explore robustness (sensitivity) of calculations. For example one may be more sceptical in one's priors than results previously observed. This more sceptical prior could be used to calculate posterior probabilities and hence sample size calculations.

The non-Bayesian methods have the advantage of being able to provide generic sample size tables for a given imprecision around the proportional response. The Bayesian approach does not provide a generic solution – each clinical trial would require specific priors and hence a specific sample size estimate. Given this however, it is recommended that Bayesian methods described here be considered for all calculations when the primary response is binary.

5. CHAPTER 5 - SAMPLE SIZE CALCULATIONS FOR CLINICAL TRIALS WITH ORDINAL DATA

This chapter will discusses the calculations for clinical trials where the expectation is that the data will take an ordinal form. The main type of ordinal data that this chapter will concentrate on, are data that is from quality of life type outcomes. The rationale for this is that quality of life (QoL) endpoints have become an increasingly important endpoint in clinical trials [de Haes and van Knippenberg, 1985; Fayers and Machin, 2000]. QoL is particularly valuable in cancer trials where an assessment of the palliative effect of treatments can be made in situations where the size of any survival advantage for a new treatment is, at most, modest.

To highlight the issues of designing a trial with a QoL outcome this chapter will use data on QoL outcome scores from a palliative clinical trial in lung cancer patients [Medical Research Council Lung Cancer Working Party, 1996]. It will be demonstrated how sub-optimal calculations can impact on sample size calculations. Finally, this chapter will describe how if one is using data from earlier trials, this impacts on sample size calculations.

In describing sample size methodologies this chapter will concentrate on methodologies that rely on the assumption of proportional odds-ratio [McCullagh, 1980; Whitehead, 1993] – using an odds-ratio to quantify a difference between treatments – and will draw heavily on the work of Julious et al [Julious, George and Campbell, 1995; Julious, George, Machin et al, 1997; Julious, Walker, Campbell et al, 2000; Campbell, Julious, Walker et al, 2000].

The chapter will be briefer than other chapters extending either the work from Normal data from Chapter 2 or the work from binary data from Chapter 4.

5.1. Aims of the Chapter

The main issues discussed in the chapter are as follows:

- To describe calculations for the estimation of sample sizes where the primary endpoint is ordinal.
- To investigate how issues such as dichotomisation can impact on sample size calculations
- To interrogate how effect sizes for parallel group studies can be generalised to cross-over trials.
- To investigate the asymptotic properties of data anticipated to be ordinal in form and how non-parametric bootstrapping methods can be used in sample size calculations.

- To assess how the conservative nature of non-inferiority and equivalence trials can impact on sample size calculations.

5.2. The Quality of Life Data

The data in this chapter are taken from a randomised parallel group controlled trial of a standard treatment against a less intensive treatment in 310 patients with small-cell lung cancer and poor prognosis [Medical Research Council Lung Cancer Working Party, 1996]. The standard treatment (*A*) consisted of a four-drug regime (etoposide, cyclophosphamide, methotrexate and vincristine) while the new less intensive treatment (*B*) under investigation contained just two of these compounds (etoposide and vincristine). The two treatment schedules were the same, comprising three cycles of chemotherapy at the same dosage. Each cycle was given on three consecutive days at three-week intervals.

The two Health Related Quality of Life (HRQoL) questionnaires used in this trial were the Hospital Anxiety and Depression Scale (HADS) [Zigmond and Snaith, 1983] and the Rotterdam Symptom Checklist (RSCL) [de Haes; van Knippenberg and Neijt, 1990].

The Hospital Anxiety and Depression Scale was developed by Zigmond and Snaith [1983]. It was designed to measure two psychological dimensions of QoL, those of anxiety and depression, in patients who were physically ill, and therefore it excluded somatic symptoms that might be due to illness. It is a self-rating questionnaire, which a patient completes in the waiting room before meeting a doctor in order to reflect how they have felt during the past week. It has 14-items which split equally into the two sub-scales and provides scores in the range 0-21 in two dimensions: anxiety and depression. Moorey, Greer, Watson et al [1991] reported that HADS is a useful instrument for measuring these dimensions in cancer patients.

The HADS has three clinically pre-defined categories for each dimension: a total score 0-7 is defined as a 'normal', 8-10 as a 'borderline-case' and 11-21 as a 'case' suggesting significant anxiety or depression.

The RSCL has two main scales, physical symptom distress and psychological distress, in addition to the scales for activity and overall evaluation. It was developed to measure the symptoms of cancer patients participating in clinical research. Patients indicate how much they have experienced particular symptoms over the last week. The RSCL psychological dimension, for example, has scores ranging from 0 to 24, where high scores constitute psychological distress. It has two clinically pre-defined categories where a total score of 0-10 is considered a 'non-case' and 11-24 is a 'case' considered to constitute psychological distress.

In the case study in this chapter setting both HRQoL questionnaires were completed together. The 310 patients' baseline scores prior to randomisation are used in this chapter

for expository purposes as the outcome for the control therapy. Two hundred and sixty six patients completed a baseline response.

Figure 5-1. Distribution of HADS anxiety scores at baseline

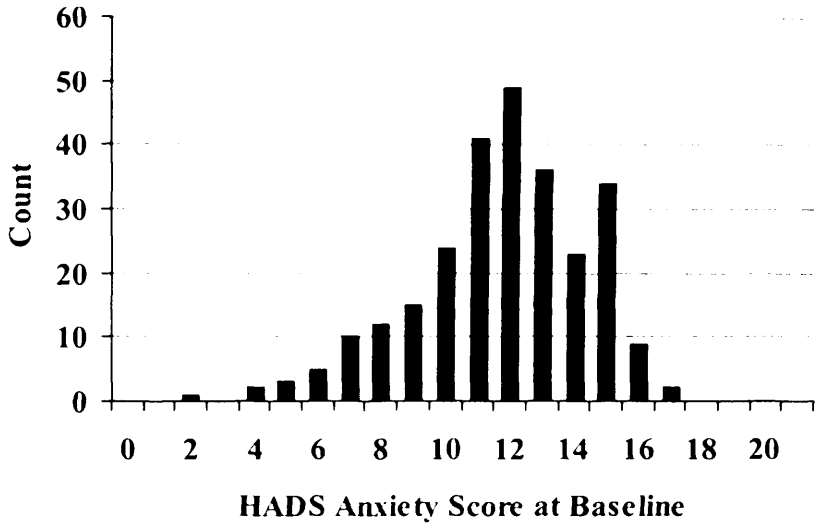
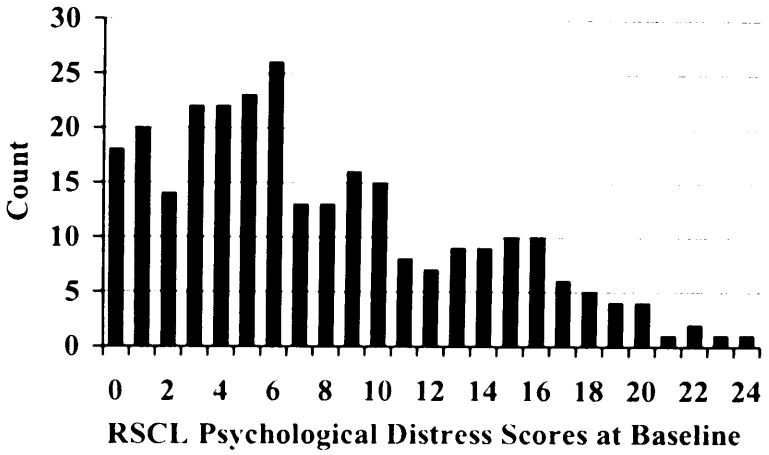


Figure 5.1 displays the distribution of the HADS anxiety scores at baseline. It is negatively skewed. Figure 5.2 shows the equivalent distribution of the RSCL psychological dimension scores. It is positively skewed. In both cases the scores do not seem to take an approximate Normal distributional form. It therefore seems that the usual mean and standard deviation are not adequate to summarise the distributions. As a consequence, in the context of this chapter, it is recommended that distribution-free techniques should be used for testing treatment differences.

Note in practice transformations such as a log transformation may be considered for such data with inference then made on the transformed scale. For the purposes of this chapter, however, transformations will not be considered.

Figure 5-2. RSCL psychological scores at baseline



5.3. Superiority Trials

5.3.1. Parallel Group Trials

5.3.1.1. Sample Sizes that are Estimated Assuming that the Population Effects are Known

Most QoL scales have categories that can be ordered, but the scores should not be treated as meaningful numbers, for example, a change in HADS from 5 to 10 is not the same as a change from 10 to 15. However, methods have been developed for sample size calculations for ordered categorical (ordinal) data [Whitehead, 1993].

As discussed in Chapter 1, in general terms for a 2-tailed, α-level test one requires the following for the variance if the test is going to have the correct power

$$\text{Var (S)} = \frac{d^2}{(Z_{1-\beta} + Z_{1-\alpha/2})^2}, \qquad (5.2.1)$$

Here d is the effect size of interest (assessed through a log-odds—ratio) with the sample variance, var(S), about the log-odds-ratio for an ordinal response estimated from [Whitehead, 1993; McCullagh, 1980; Jones and Whitehead, 1979, Campbell, Julious and Altman, 1995]

$$Var\,(S) = \frac{6}{n\left(1 - \sum_{i=1}^{k} \bar{p}_i^3\right)}, \qquad (5.2.2)$$

Here k is the number of categories on the QoL instrument, $\bar{p}_i$ is the mean proportion expected in category i, that is, $\bar{p}_i = (p_{Ai} + p_{Bi})/2$, where p_{Ai} and p_{Bi} are the proportions

anticipated in category i for the two treatment groups A and B respectively and α and β are the overall type I and type II errors respectively with $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ denoting the percentage points of a standard Normal distribution for these two errors. Here n is the sample size per group. Note that in this chapter, as for binary data, the issue of allocation ratios between treatments will be ignored.

Now by equating (5.2.1) with (5.2.2) one has [Julious, George and Campbell, 1995; Julious, George, Machin et al, 1997; Julious, Walker, Campbell et al.2000; Campbell, Julious, Walker et al, 2000; Whitehead, 1993; Dark, Bolland and Whitehead, 2003; Lee, Song, Kang et al, 2002; Rabbee, Coull, Mehta et al, 2003; Campbell, Julious and Altman, 1995]

$$n = \frac{6\left[Z_{1-\beta} + Z_{1-\alpha/2}\right]^2 / (\log OR)^2}{\left[1 - \sum_{i=1}^{k} \bar{p}_i^3\right]} \quad . \qquad (5.2.3)$$

Equation (5.2.1) is based on the Mann-Whitney U-test for ordered categorical data. It estimates the sample size based on the odds ratio (OR) of a patient being in a given category or less in one treatment group compared to the other group.

Note that a form of this equation was used in Chapter 4 for binary data - a binary response being a special case of (5.2.3). For an analysis under the assumption of proportion odds the anticipated effect size is expressed as an odds ratio defined as:

$$OR = \frac{p_{Ai}(1 - p_{Bi})}{p_{Bi}(1 - p_{Ai})}.$$

This is a measure that is not immediately straightforward to interpret for binary data and, as a consequence, is more difficult for an ordinal response. It is best to discuss the application of (5.2.3) through a worked example.

There are alternative ordinal methodologies [Hilton and Mehta, 1993; Lachin, 1977; Noether, 1987] for sample size calculation for ordinal data. Indeed the methodology of Noether [1987] is of particular interest as it is written simply in terms of the probability of one group being greater than the other. However, this chapter will concentrate on (5.2.3) as it has the advantage of being thought of as generalised from binary methodology and as a result (as will be highlighted later) has a measure of effect that can be interpreted in terms of the binary response.

5.3.1.2. Worked Example

5.3.1.3. Full Ordinal Scale

When designing a clinical trial to estimate the odds ratio one can utilise the predefined clinical cut points that the HADS and RSCL each provide. For example, 27.1% of patients at baseline are defined as borderline cases or better on the HADS Anxiety dimension score at baseline (see Table 5.1), that is, 27.1% record values resulting in a score of ≤10. This one could take, for expository purposes, as what one would expect on standard therapy (*S*). The odds with *S* is thus 0.271/(1-0.271)=0.372. Suppose a new therapy (*T*) is to be studied and the investigator decided that a clinically meaning effect is one that would increase the proportion of non-cases to 40.0% or a postulated odds of 0.40/(1-0.40)=0.67. The ratio of these odds gives OR=0.372/0.667=0.56 in favour of *T*. This value can then be used as the basis for the sample size calculation.

Equation (5.2.3) makes no assumption about the distribution of the data, but it does assume proportional odds between the treatments across the QoL dimension. This implies that the odds ratios are identical for each pair of adjacent categories throughout the scale. What this means practically can be highlighted by extending the example given above. When using the pre-defined clinical cut point for 'non-cases' the investigator anticipated the OR would be 0.56. The assumption of proportional odds implies that, if instead of using ≤10 as the definition of a 'non-case', ≤9 had been used, one would nevertheless obtain OR_9=0.56, and so on for OR_8, OR_7 etc. Thus, although the actual observed odds ratios might differ from each other across the scale, the corresponding population values are all equal which implies that $OR_1=OR_2=OR_3=\ldots=OR_{21}=0.56$. However, the calculations of sample size using (5.2.3) are robust to departures from this ideal, provided all the odds ratios indicate an advantage to the same treatment [Julious, George, Machin et al, 1995; Julious, Walker, Campbell et al, 2000].

Table 5-1. Frequency of responses on the HADS anxiety scores as baseline for patients with small-cell lung cancer

Category	Score	Number of Patients
Normal	0	0
	1	0
	2	1
	3	0
	4	2
	5	3
	6	5
	7	10
Borderline	8	12
	9	15
	10	24
Clinical Case	11	41
	12	49
	13	36
	14	23
	15	34
	16	9
	17	2
	18	0
	19	0
	20	0
	21	0
	Total	266
Normal	0 – 8	21 (7.9%)
Borderline	9 – 10	51 (19.2%)
Clinical Case	11-21	194 (72.9%)
Mean		11.70
SD (σ)		2.66
Median		12

Using the odds-ratio of 0.56 the anticipated new therapy responses can be derived as per Table 5.2. From these anticipated responses an estimate of the variance can be made from (5.2.2) which when placed in (5.2.3) with the odds-ratio gives an estimate of the sample size of 188 patients per arm (for 90% power and two sided Type I error rate of 5%).

Table 5-2. Anticipated percentages of response on the HADS anxiety scores for standard treatment (S) and new treatment for patients with small-cell lung cancer

Category	Score*	Standard Therapy (S)		New Therapy (T)	
		Percentage (p_{Si})	Cumulative percentage (Q_{Si})	Percentage (p_{Ti})	Cumulative percentage (Q_{Ti})
Normal	0 - 3	0.4	0.4	0.7	0.7
	4	0.8	1.2	1.4	2.1
	5	1.1	2.3	1.9	4.1
	6	1.9	4.2	3.2	7.3
	7	3.8	8.0	6.2	13.5
Borderline	8	4.5	12.5	6.9	20.4
	9	5.6	18.1	8.0	28.4
	10	9.0	27.1	11.6	40.0
Clinical Case	11	15.4	42.5	17.0	57.0
	12	18.4	60.9	16.6	73.6
	13	13.5	74.4	10.3	83.9
	14	8.6	83.0	5.8	89.8
	15	12.8	95.8	7.8	97.6
	16	3.4	99.2	1.9	99.6
	17 – 21	0.8	100.0	0.4	100.0

*Note The 22 categories of Table 1 are reduced to $k = 15$.

The application of proportional odds therefore allows that, if the distribution of one of the treatment groups can be specified, then the anticipated cumulative proportions for the other treatment can be directly derived. Hence, with prior knowledge of the distribution of just one treatment group and an anticipated *OR*, obtained about any cut point on the QoL scale, an estimate of the sample size can be obtained.

5.3.1.4. Effects of Dichotomisation

An advantage of the HADS and RSCL instruments for the process of anticipating the effect size and consequent sample size is that they both have predefined definitions of what constitutes a 'case' and which can then be used to obtain a value of a readily interpretable effect size. This effect size, here expressed as an odds ratio, can thus be extended across the full QoL scale and an estimate of the sample size made.

These cut-offs, however, can encourage some researchers to dichotomise QoL scales to calculate sample sizes. For example, with the HADS Anxiety dimension, one of the cut-offs can classify subjects either as a 'clinical case' or 'borderline' or better. For this, now binary, situation (5.2.3) can still be used to estimate a sample size but ignoring the full ordered categorical nature of the data, may result in a substantial over-estimation of the

sample size. For example, if a clinically meaningful difference was set again at 0.56 around the cut off of non-cases/clinical cases on the HADS Anxiety score then by dichotomising (5.2.3) gives an estimate of the sample size of 277 compared to only 188 when all $k = 22$ categories are used in the calculations. This is a potential over-estimate of 47% in the necessary sample size if the data was analysed using all 22 categories.

Obviously, if the intention is to analyse the scale as a dichotomous endpoint then the sample size may be appropriate – although this approach may be questioned also as wasting patients.

5.3.1.5. Effects of Additional Points

It may not be necessary to use the full categorical scale. For example, with HADS there is an additional category of 'normal' for subjects with a score of ≤8 and just under 8% of patients are classified as such on the anxiety dimension. If one then calculated the sample size using the k=3 groups of 'normal', 'borderline-case' and 'clinical-case' as the categories, the estimated sample size, from (5.2.3), is N_3=267 subjects - only a marginally closer estimate. However, if one identified an additional category of 'severe-clinical-case' for subjects with a HADS score ≥14 [Julious, George, Machin et al, 1997; Julious, Walker, Campbell et al, 2000] and based the sample size calculations on the 4 categories, the estimated sample size is of 210 patients is now quite close to the optimal 188.

Dichotomising the QoL scale in order to estimate a sample size, and consequently analysing the data as ordinal, should be avoided if possible as sample sizes could be unnecessarily inflated. However, knowledge of anticipated responses in only a handful of categories can give sample size estimates that are more precise for only a modest increase in the complexity of the calculations.

Table 5-3. Correction factor to be used when the number of categories is less than 5

Number of Categories	Mean Proportions Anticipated	Correction Factor
2	$\bar{p}_1 = \bar{p}_2 = 0.5$	1.333
3	$\bar{p}_1 = \bar{p}_2 = \bar{p}_3 = 0.333$	1.125
4	$\bar{p}_1 = \bar{p}_2 = \bar{p}_3 = \bar{p}_4 = 0.25$	1.067
5	$\bar{p}_1 = \bar{p}_2 = \bar{p}_3 = \bar{p}_4 = \bar{p}_5 = 0.2$	1.042

The reason why ignoring the ordinal scale substantially increases the sample size is due to the increase in variance estimated from (5.2.2). Table 5.3 illustrates this point. The minimum the variance can be for any number of categories on a scale is for the special case

where the anticipated mean responses for each category are equal i.e. $\bar{p}_1 = \bar{p}_2 = \bar{p}_3 = ...\bar{p}_{k-1} = \bar{p}_k$. If this result is placed into (5.1.1) one can obtain the anticipated relative variances for different numbers of categories for the most optimal responses. The ratio of these variances can in turn give inflation factors for the sample size for different numbers of categories relative to the optimum number of categories i.e. a continuous scale where $1 - \sum_{i=1}^{k} \bar{p}_i^3 = 1$.

From Table 5.3 one can see that for the optimum mean responses one would anticipate a 33% increase compared to a continuous response as opposed to just 5% for 5 categories [Campbell, Julious and Altman, 1995]. Thus, what these results show is that dichotomising could lead to a serious inflation of the sample size. Even using only a little extra information (from extra categories) can substantially improve a sample size estimate.

Note also that another common way to calculate sample sizes for ordinal responses is to use the Normal data methodology described in Chapter 2. Julious et al demonstrated that assuming the data takes a Normal form might also give misleading results [Julious, George and Campbell, 1995; Julious, Walker, Campbell et al, 2000]. This is mainly due to the fact that the distributional form of an ordinal response may be asymmetric. This would lead to asymmetric sample size estimates depending on the sign of the effect of interest. In comparison, however, for Normal data a mean shift in either direction would give the same size due the symmetric form of the data.

5.3.1.6. Sensitivity Analysis about the Estimates of the Population Effects Used in the Sample Size Calculations

5.3.1.7. Extending the Results from Normal Data

In Chapter 2 it was described how, for data that take a Normal form, the sensitivity of a trial design to the variance could be investigated using the degrees of freedom of the variance estimate used in the calculations. Using the degrees of freedom for the variance, and the chi-squared distribution, the sensitivity of the study can be investigated for high plausible values for the variance, taken from the upper 95th percentile for the variance using the following formula

$$s^2(95) < \frac{df}{\chi^2_{0.95,df}} s^2. \qquad (5.2.4)$$

Where s^2 is a sample variance from the trial(s) being used for planning purposes. To generalise this result to ordinal data (5.2.2) could be used in (5.2.4) to obtain an upper estimate of the variance.

Next one can rewrite (5.2.3) to be written in terms of power i.e.

$$1-\beta=\Phi\left(\sqrt{n\left[1-\sum_{i=1}^{k}\bar{p}_i^3\right](\log OR)^2/6}-Z_{1-\alpha/2}\right). \qquad (5.2.5)$$

With this result an upper estimate of the variance could be used to assess the sensitivity of the study to assumptions around the sample variance being used in the planning. However, for binary data discussed in Chapters 3 and 4 the equivalent result to (5.2.4) does not hold, as the variance for these data does not follow a chi-squared distribution. The assumptions around the result (5.2.4) will now be interrogated for ordinal data.

5.3.1.8. Simulation Investigation of the Ordinal Variance

Ordinal data to an extent fall between binary data and Normal data. Thus, to investigate the appropriateness of using (5.24) with (5.25) a simulation was undertaken for different sample sizes, number of categories (of the ordinal response) and expected mean responses. It is not anticipated that the variance for ordinal data (described in the context of this chapter) will follow a chi-squared distribution, as this requires Normality and hence the simulations are undertaken for interest. The simulation was done in SAS [1990]. For investigation 10,000 simulations were undertaken.

For each simulation the ratio of the 'sample' variance over the 'population' variance was calculated which would be plausibly expected approximated to follow a chi-squared distribution i.e.:

$$\frac{(n-1)s^2}{\sigma^2}=(n-1)\frac{6n/\left[1-\sum_{i=1}^{k}\bar{p}_i^3\right]}{6n/\left[1-\sum_{i=1}^{k}\bar{\pi}_i^3\right]}=(n-1)\frac{1/\left[1-\sum_{i=1}^{k}\bar{p}_i^3\right]}{1/\left[1-\sum_{i=1}^{k}\bar{\pi}_i^3\right]},$$

where $\bar{p}_i$ are the estimated mean responses from the simulation and $\bar{\pi}_i$ are the population mean responses from which each simulation was drawn.

Figures 5.3 to 5.5 give some indicative simulations for 3, 4 and 5 category responses. One can see from these simulations that the approximation to the chi-squared gets better the more categories one has and also the greater the sample size. Overall, however, the simulations indicate a reasonable approximation to the chi-squared distribution in this instance, although not great.

Figure 5-3. Chi- probability plots for a 3 category variable for anticipated responses of 0.4, 0.3 and 0.3 for the 3 categories

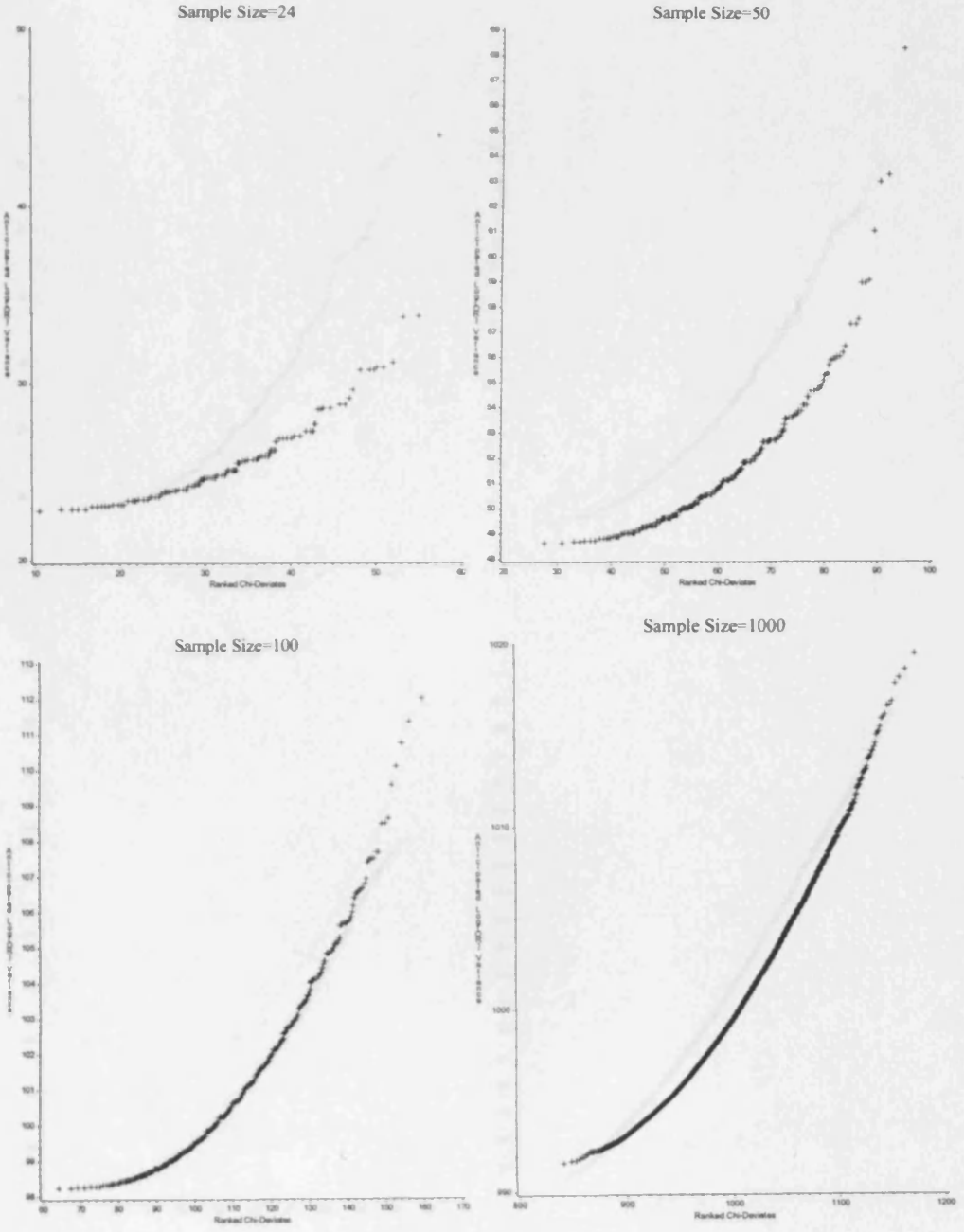


Figure 5-4. Chi- probability plots for a 4 category variable for anticipated responses of 0.3, 0.3, 0.2 and 0.2 for the 4 categories

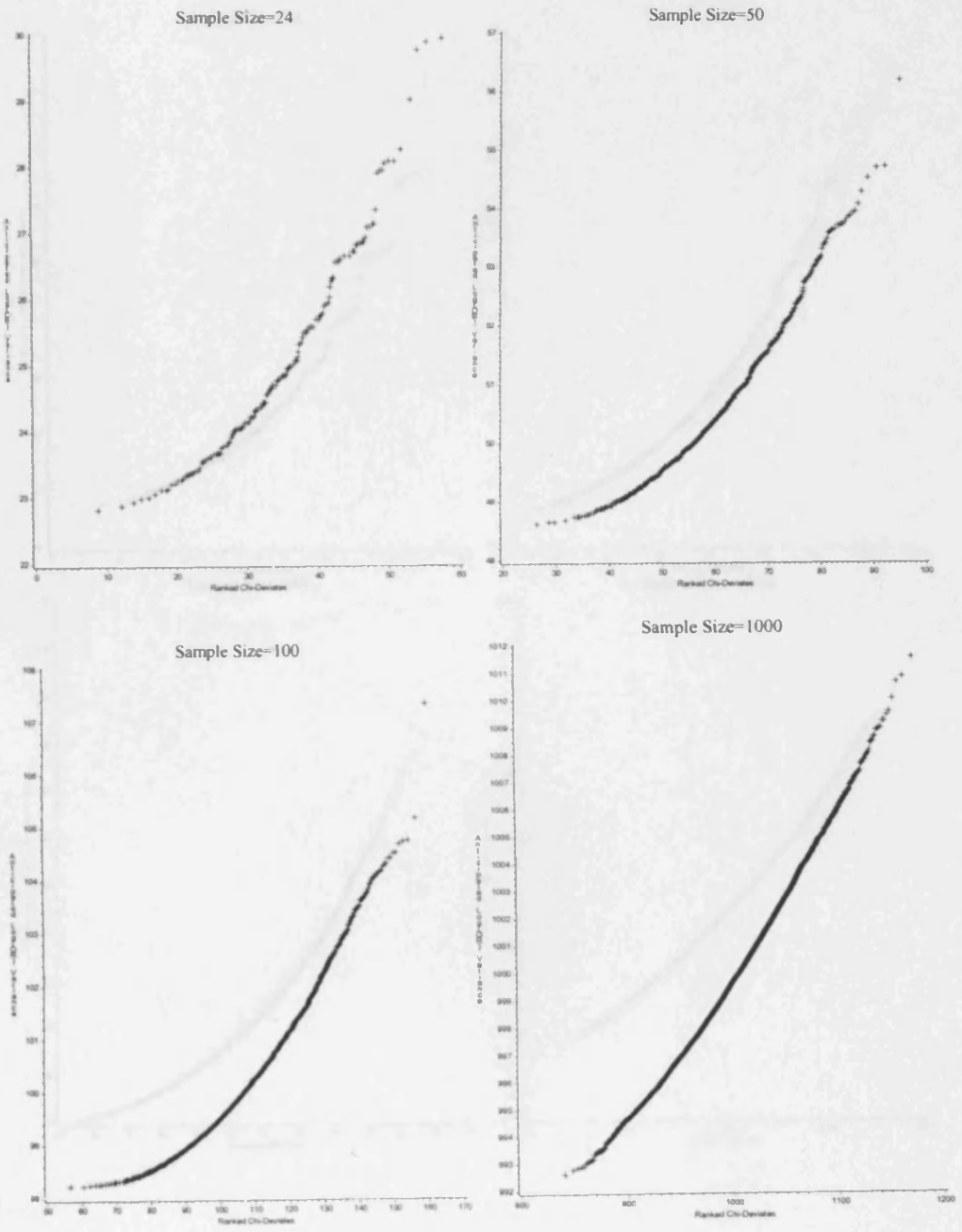
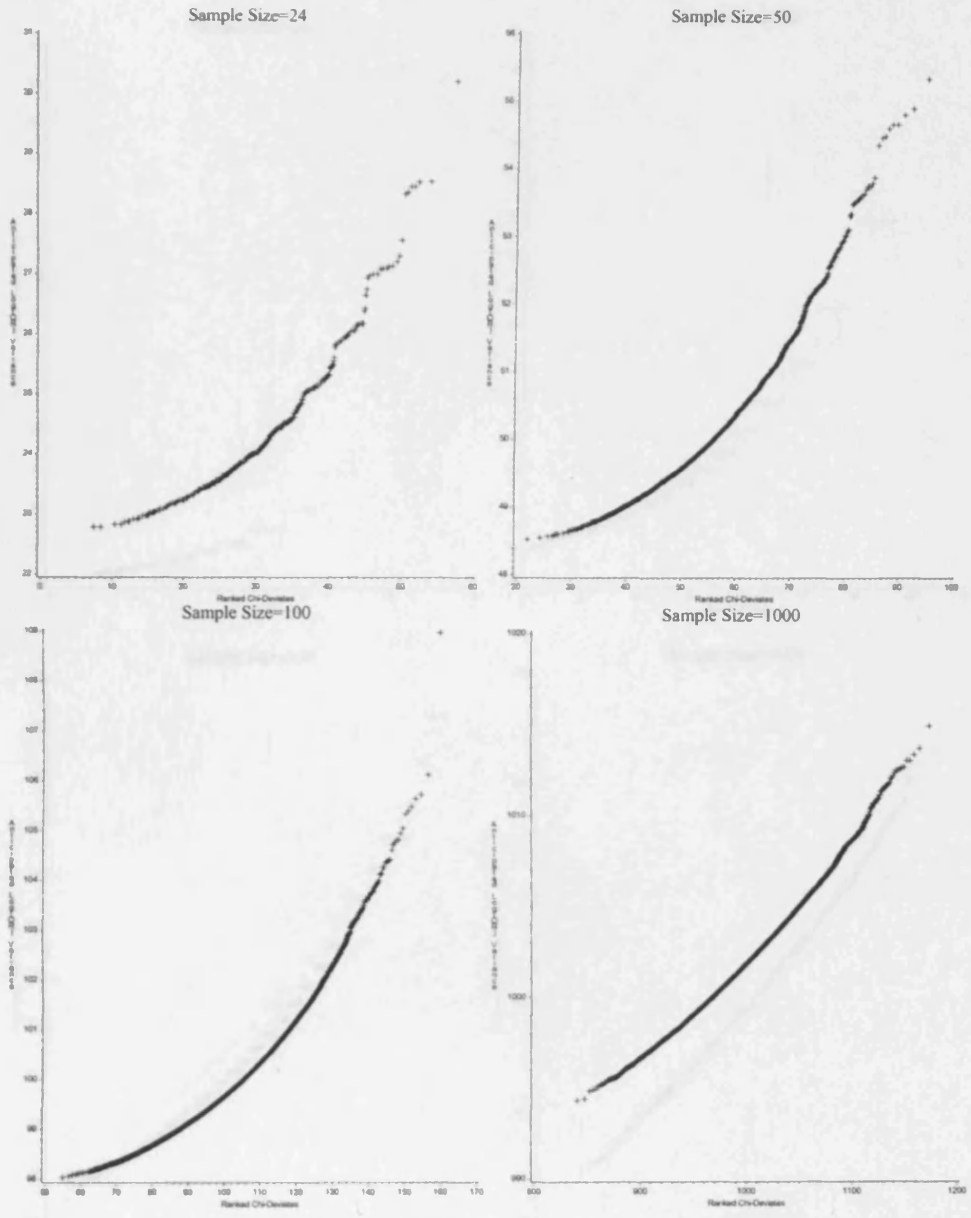


Figure 5-5. Chi- probability plots for a 5 category variable for anticipated responses of 0.3, 0.2, 0.2, 0.15 and 0.15 for the 5 categories



Figures 5.6 to 5.8 give alternative simulations (again for 3, 4 and 5 category responses) but here one category is expected to dominate. In this instance, except for large sample sizes, the approximation to the chi-squared is weak. These results are, as one would anticipate for with one category dominating the data, similar in parametric form to that of a binary response. As discussed in Chapters 3 and 4 the variance for binary responses does not follow a chi-squared distribution.

Figure 5-6. Chi- probability plots for a 3 category variable for anticipated responses of 0.8, 0.1 and 0.1 for the 3 categories

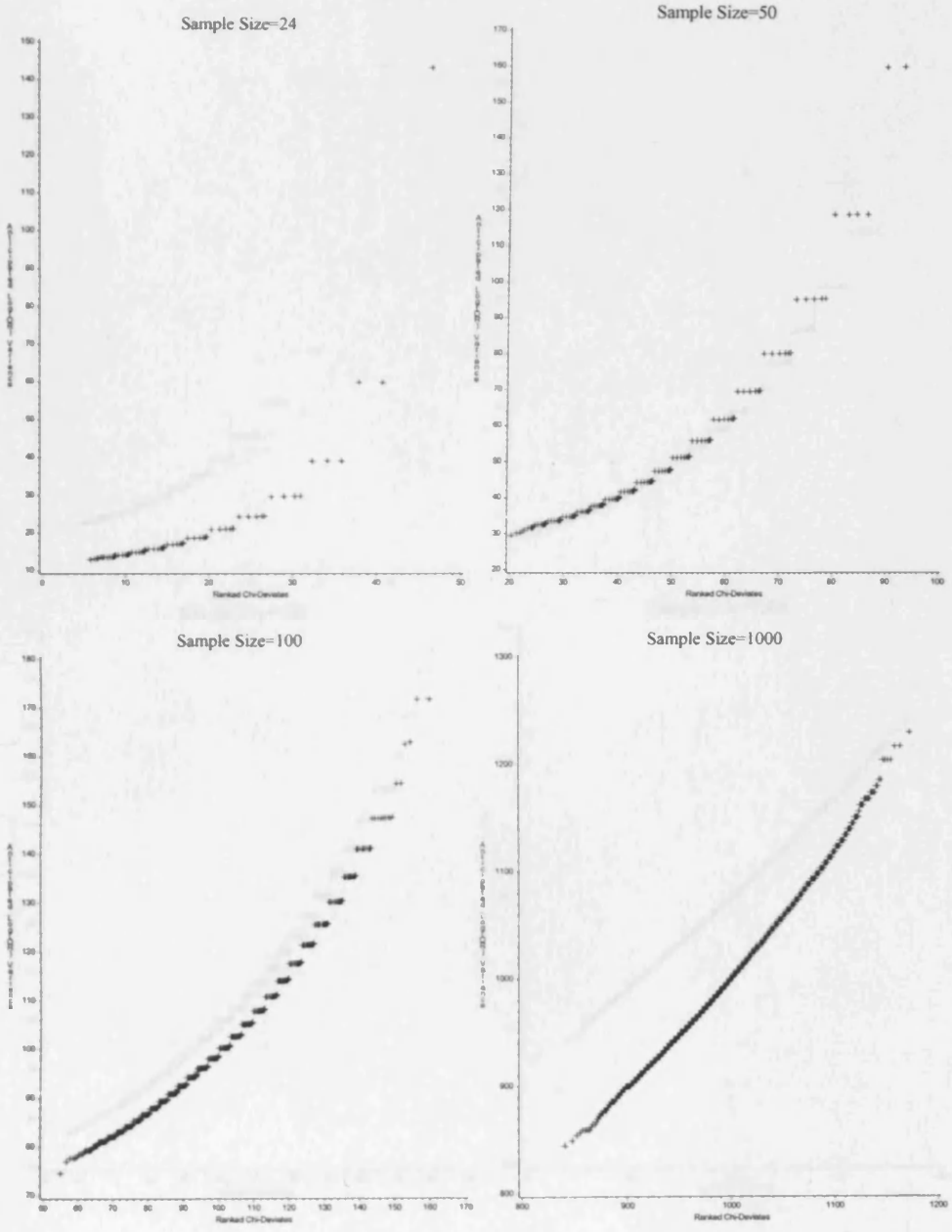


Figure 5-7. Chi- probability plots for a 4 category variable for anticipated responses of 0.7, 0.1, 0.1 and 0.1 for the 4 categories

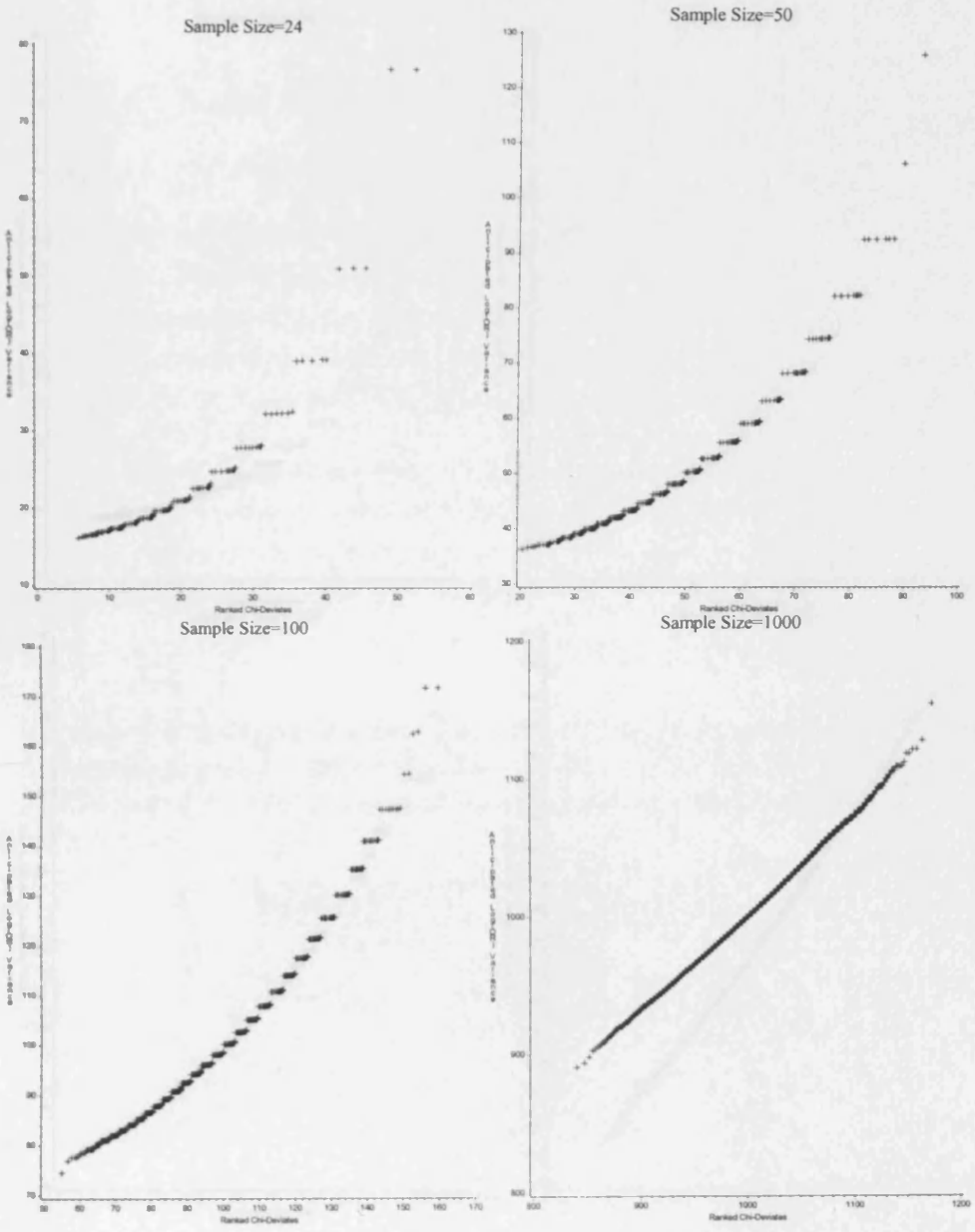
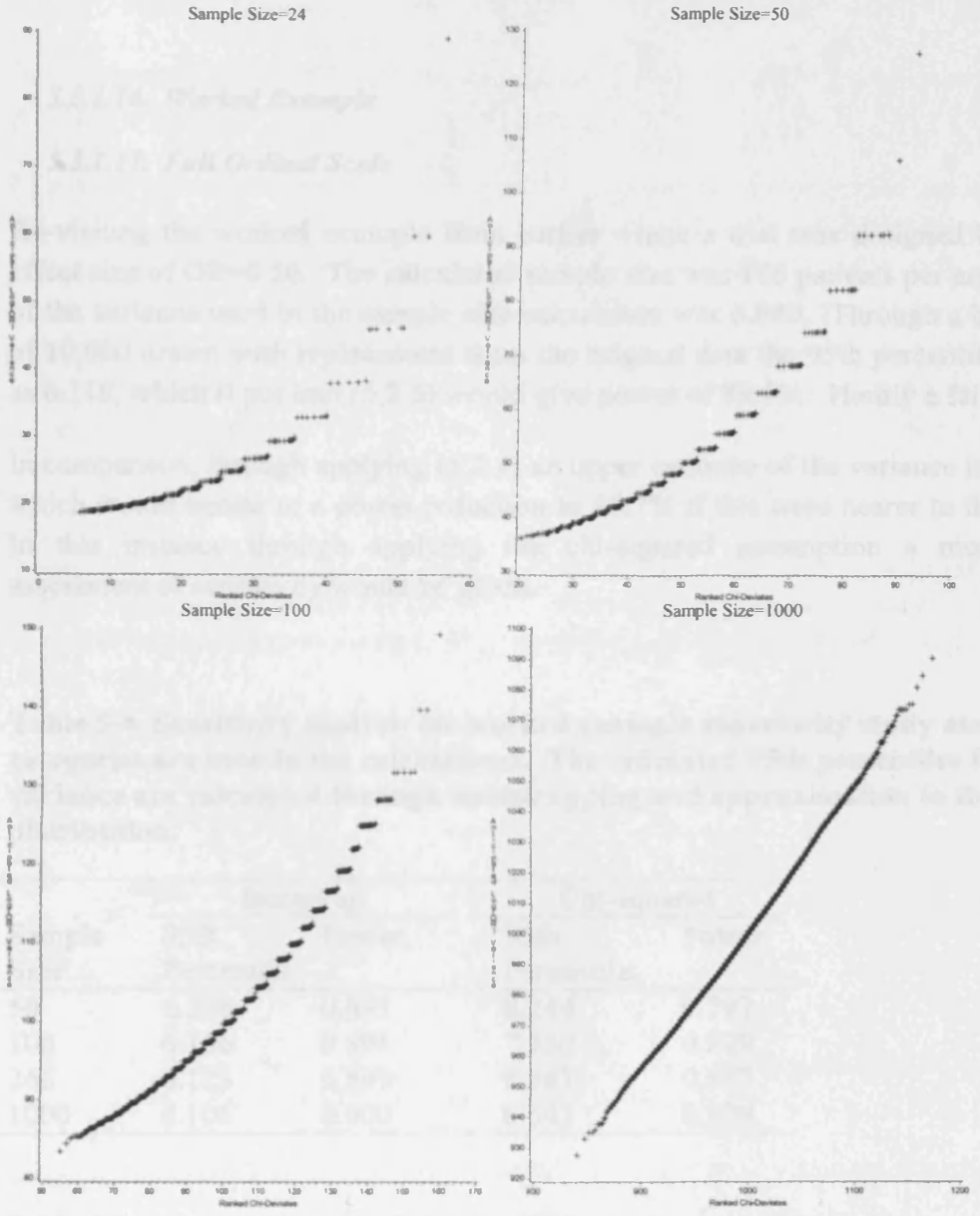


Figure 5-8. Chi- probability plots for a 5 category variable for anticipated responses of 0.6, 0.1, 0.1, 0.1 and 0.1 for the 5 categories



5.3.1.9. Bootstrapping

What the results illustrate in this sub section is that the variance for an ordinal response does not follow a chi-squared distribution (although other distributional forms can not be ruled out. A number of authors have described how non-parametric bootstrapping can be applied [Efron and Tibshirani; Hall, 1993; Julious, 2001; Keene, 2002]. These arguments can be extended to the situation here. Hence, a solution to the problem is to form a bootstrap distribution for the sample variance for a particular example and from this take a

95th percentile. This one tailed bootstrap upper percentile could be used to investigate the sensitivity of a study. An example is the best way to illustrate the solution.

5.3.1.10. Worked Example

5.3.1.11. Full Ordinal Scale

Re-visiting the worked example from earlier where a trial was designed based around an effect size of OR=0.56. The calculated sample size was 188 patients per arm. The estimate of the variance used in the sample size calculation was 6.089. Through a bootstrap sample of 10,000 drawn with replacement from the original data the 95th percentile was estimated as 6.119, which if put into (5.2.5) would give power of 89.9%. Hardly a fall at all.

In comparison, through applying (5.2.4) an upper estimate of the variance is given as 6.983, which would equate to a power reduction to 85.7% if this were nearer to the true variance. In this instance through applying the chi-squared assumption a more conservative assessment of sensitivity would be given.

Table 5-4. Sensitivity analysis for worked example superiority study assuming all categories are used in the calculations. The estimated 95th percentiles for the variance are calculated through bootstrapping and approximation to the chi-squared distribution.

	Bootstrap		Chi-squared	
Sample Size	95th Percentile	Power	95th Percentile	Power
50	6.296	0.891	8.244	0.797
100	6.156	0.898	7.580	0.829
266	6.125	0.899	6.983	0.857
1000	6.101	0.900	6.545	0.879

Table 5.4 gives a summary of the sensitivity assessment along with repeated calculations assuming the same distribution of responses was observed but drawn from sample sizes of 50, 100, 266 (the actual sample size) and 1000. For each case assuming a chi-squared distribution gives a more conservative assessment.

5.3.1.12. Four Point Scale

In the same example earlier for the same effect size the 22 point scale was reduced to 4 (using clinical cut-offs) for ease of calculations. The estimated variance as a result was

increased to 6.796, which as a result increased the sample size to 210 patients. A bootstrap 95th percentile for the variance is estimated as 7.041 which if was nearer the true variance would mean the powered reduced to 89.1%.

The corresponding variance estimate from (5.2.4) is 7.795 would mean a reduction in power to 85.8%.

Table 5.5 gives a summary of the sensitivity assessment along with repeated calculations for different sample sizes. A couple of points are worth highlighting from this table. The first is that although the bootstrap estimates are now nearer those estimated from a chi-squared, the chi-squared estimates are still a little conservation. The second point is that although, in terms of the initial sample size calculation, discarding categories does not have a major effect, comparing Table 5.5 with Table 5.4 it does seem in this instance that one's calculations do become more sensitive to the assumptions about the variance.

Table 5-5. Sensitivity analysis for worked example superiority study assuming 4 categories are used in the calculations. The estimated 95th percentiles for the variance are calculated through bootstrapping and approximation to the chi-squared distribution.

	Bootstrap		Chi-squared	
Sample Size	95th Percentile	Power	95th Percentile	Power
50	8.522	0.826	9.202	0.797
100	7.251	0.882	8.461	0.829
266	7.041	0.891	7.795	0.857
1000	6.911	0.896	7.206	0.879

5.3.1.13. Calculations Taking Accounting of the Imprecision of the Estimates of the Population Effects Used in the Sample Size Calculations

If the results from Chapter 2 could be extended to ordinal responses then to account for the imprecision in the variability the following result could be applied

$$n = \frac{6\left[tinv\left(1-\beta, df, Z_{1-\alpha/2}\right)\right]^2 / \left(\log OR\right)^2}{\left[1 - \sum_{i=1}^{k} p_i^3\right]}, \qquad (5.2.6)$$

where $TINV(\bullet,m,a)$ denotes the (monotonically increasing) inverse function of the cumulative distribution of a Students non-central t distribution with m degrees of freedom and non-centrality parameter a. The degrees of freedom (d.f.) in the formula refers to the degrees of freedom about the variance estimate used in the sample size calculation.

In Chapters 3 and 4 it was highlighted how an equation of the form of (5.2.6) could not be applied to binary data as the assumption of the variance following a chi-squared distribution does not hold. It was recommended instead that an equation of the form

$$1-\beta = \frac{1}{0.998}\sum_{perc=0.001}^{0.998} 0.5\left[\begin{array}{l}\Phi\left(\sqrt{n}(\log(OR))^2/[\text{var}(\log(OR))]_{perc} - Z_{1-\alpha/2}\right) \\ +\Phi\left(\sqrt{n}(\log(OR))^2/[\text{var}(\log(OR))]_{perc} - Z_{1-\alpha/2}\right)\end{array}\right], \qquad (5.2.7)$$

be applied and the sample size be estimated through numerical methods. Remember that for binary data values for $[\text{var}(\log(OR))]_{perc}$ were estimated through the percentiles from the control prevalence – from which the variance and sample size were base.

For ordinal data there seems also to be a similar issue with the variance about the log(OR) also seeming not to following a chi-squared distribution as a general rule. To assess sensitivity it was recommended that a bootstrap distribution be built around the variance and a 95th percentile taken from this. It is now recommended that the same arguments be extended to provide values for $[\text{var}(\log(OR))]_{perc}$ to be put into (5.2.7). To do this, follow the following steps

1. Generate an empirical bootstrap distribution for $[\text{var}(\log(OR))]_{perc}$ through sampling with replacement from the original distribution.
2. Rank the empirical distribution of $[\text{var}(\log(OR))]_{perc}$ in order of size.
3. Take the smallest value as the 1st percentile, 2nd smallest as the 2nd percentile etc
4. Use these empirical percentiles in (5.2.7) and calculate the average power across these for a given sample size.
5. Iterate the sample size to the required power is reached.

It is again best to highlight the calculations through worked example.

5.3.1.14. Worked Example

5.3.1.15. Full Ordinal Scale

Remember the worked example from earlier where a trial was designed with a calculated sample size of 188 patients per arm. The variance of 6.089 was estimated from a trial of 266 evaluable patients. Forming an empirical bootstrap distribution of 10,000 drawn with replacement from the original data for the percentiles for the variance (5.2.7) also gives 188 patients per arm.

In comparison using (5.2.6) gives a sample size of 190 per arm - a slight increase in the required sample size.

Table 5.6 gives a summary of the sample size calculations along with repeated calculations assuming the same distribution of responses was observed but drawn from sample sizes of 10, 15, 50, 100 and 266 (the actual sample size). For each case assuming (5.2.6) gives a more conservative sample size.

Table 5-6. Sample sizes for worked example superiority study assuming all categories are used in the calculations. The sample sizes were estimated taking percentiles for the variance calculated through bootstrapping and approximation to a non-central t-distribution.

	Calculated Sample Size	
Original Sample Size	Bootstrap	Non-central t
10	196	251
25	191	209
50	189	198
100	189	193
266	188	190

5.3.1.16. Four Point Scale

Reducing the scale to a 4 point one increases the sample size estimate to 210 patients. Taking into account the original variance was estimated from 266 patients, through bootstrapping, and (5.2.7) also gives a sample size of 210. In comparison (5.2.6) gives a sample size of 212.

Table 5-7. Sample sizes for worked example superiority study assuming 4 categories are used in the calculations. The sample sizes were estimated taking percentiles for the variance calculated through bootstrapping and through approximation to the chi-squared distribution.

	Calculated Sample Size	
Original Sample Size	Bootstrap	Chi-Squared
10	234	281
25	218	234
50	214	224
100	212	215
266	210	212

Table 5.7 gives a summary of the sample calculations along with repeated calculations assuming the bootstrap sample was taken from different sample sizes. It is worth noting that in comparison to Table 5.6 that imprecision of the variance estimate (assessed through the original sample size the estimate was drawn from) has greater effect on the 4 point scale.

When using all the categories, accounting for the imprecision has little effect on the sample size calculations (for the case study described) and so this chapter will now concentrate on the 4-point scale.

5.3.1.17. Calculations Taking Accounting of the Imprecision of the Estimates of the Population Effects Used in the Sample Size Calculations – Bayesian Methods

As the effect size (with respect to the odds-ratio) is fixed the binary methods described in Chapter 4 could be extended to ordinal data when there is a cut off. If there is a prior belief, expressed by the mode and a percentile, of the response around the cut off for a control treatment arm. From previously observed trial data and the fixed odds-ratio a posterior response distribution can be built for the investigative arm. This could be used in (5.2.7) to estimate sample sizes.

Alternatively the work with beta distribution could be generalised to a multinomial direchlet distribution, which could be used for priors [Ferguson 1973, 1974; Escobar 1994, 1995].

5.3.2. Cross-over Trials

5.3.2.1. Sample Sizes that are Estimated Assuming that the Population Effects are Known

Remember from Chapter 4 how the methodology for parallel groups could be generalised to that for cross-over trials when the data is binary. The practical consequence was that equivalent effect sizes could be used and the sample size for one arm of a parallel group trial could be taken as the total sample size for a cross-over.

Table 5-8. Summary table of hypothetical cross-over trial

		Treatment B				
		1	2	3	4	
Treatment A	1	λ_{11}	λ_{12}	λ_{13}	λ_{14}	P_{A1}
	2	λ_{21}	λ_{22}	λ_{23}	λ_{24}	P_{A2}
	3	λ_{31}	λ_{32}	λ_{33}	λ_{34}	P_{A3}
	4	λ_{41}	λ_{42}	λ_{43}	λ_{44}	P_{A4}
		P_{B1}	P_{B2}	P_{B3}	P_{B4}	1

The same principals as applied to binary data can be extended to ordinal data through applying the results of Agresti [1993, 1999]. Table 5.8 gives a table of hypothetical cross-over data where each cell of the 4x4 table is derived from the marginal totals.

Table 5.9 gives the 2x2 tables around each cut off on the ordinal scale corresponding to Table 5.8. Under the assumption of proportional odds the odds ratios from each of these tables should equal each other. Also extending the work from Chapter 4 the odds-ratios from each of these tables will also approximately equal the equivalent odds-ratios calculated from the marginal totals i.e. those expected from a parallel group trial.

Table 5-9. Summary table of hypothetical cross-over trial revisited

a. First cut off

		Treatment B		
		1	2+3+4	
Treatment A	1	p_{11}	$p_{12}+p_{13}+p_{14}$	Q_{A1}
	2+3+4	$p_{21}+p_{31}+p_{41}$	$p_{22}+p_{23}+p_{24}+p_{32}+p_{43}+p_{44}+p_{42}+p_{43}+p_{44}$	$1-Q_{A1}$
		Q_{B1}	$1-Q_{B2}$	1

b. Second cut off

		Treatment B		
		1+2	3+4	
Treatment A	1+2	$p_{11}+p_{12}+p_{21}+p_{22}$	$p_{13}+p_{14}+p_{23}+p_{24}$	Q_{A2}
	3+4	$p_{31}+p_{32}+p_{41}+p_{42}$	$p_{33}+p_{34}+p_{43}+p_{44}$	$1-Q_{A2}$
		P_B	$1-P_B$	1

b. Third cut off

		Treatment B		
		1+2+3	4	
Treatment A	1+2+3	$p_{11}+p_{12}+p_{13}+p_{21}+p_{22}+p_{23}+p_{31}+p_{32}+p_{33}$	$p_{41}+p_{42}+p_{43}$	Q_{A3}
	4	$p_{41}+p_{42}+p_{43}$	p_{44}	$1-Q_{A3}$
		Q_{B3}	$1-Q_{B3}$	1

To obtain an overall estimate of the odds-ratio Agresti [1993, 1999] gave the following result

$$OR = \frac{\sum_{i<j}(j-i)\lambda_{ij}}{\sum_{j<i}(i-j)\lambda_{ij}}, \qquad (5.2.8)$$

where and i and j are the row and column numbers respectively and λ_{ij} are the cell counts corresponding to the cell counts (see Table 5.8). The variance for (5.2.8) is defined as

$$\text{var}[\log(OR)] = \left(\frac{\sum_{i<j}(j-i)^2\lambda_{ij}}{\left[\sum_{i<j}(i-j)\lambda_{ij}\right]^2} + \frac{\sum_{i>j}(j-i)^2\lambda_{ij}}{\left[\sum_{i>j}(j-i)\lambda_{ij}\right]^2} \right), \qquad (5.2.9)$$

which can be rewritten in terms of the cell probabilities, p_{ij}, as follows

$$\text{var}[\log(OR)] = \frac{1}{n}\left(\frac{\sum_{i<j}(j-i)^2 p_{ij}}{\left[\sum_{i<j}(i-j)p_{ij}\right]^2} + \frac{\sum_{i>j}(j-i)^2 p_{ij}}{\left[\sum_{i>j}(j-i)p_{ij}\right]^2} \right). \qquad (5.2.10$$

By definition this odds-ratio would equate to that one would expect from a parallel group study which is a useful result. To calculate the required sample size one could equate (5.2.8) and (5.2.10) with (5.2.1) to give a sample size estimate for the total sample size of the form

$$n = \frac{\left[Z_{1-\beta} + Z_{1-\alpha/2}\right]^2 \text{var}(\log(OR))}{\left[\log(OR)\right]^2}. \qquad (5.2.11)$$

Again it is best to highlight the calculations through a worked example

Note to undertake such a period-adjusted analysis similar to described for binary data in Chapter 4 (5.2.8) would need to be applied to each sequence and the average (on the log scale) taken to obtain a period adjusted effect. However, generalising the results for binary data, whilst period-adjusted analyses are important it is not an important issue when designing a trial.

5.3.2.2. Worked Example

Suppose an investigator wishes to design a trial where the outcome is a four point ordinal response. Whilst the anticipated responses on the control treatment (treatment A) are given in the final "overall" column of Table 5.10. The effect size of interest is an odds-ratio of 0.56. From this odds-ratio the anticipated responses for the investigative treatment are given the final row – assuming proportional odds of the marginal responses. The Type I and Type II errors are set at 5% and 10%.

Table 5-10. Summary table of cross-over trial for worked example

		Treatment B				
		1	2	3	4	Overall
Treatment A	1	0.011	0.021	0.035	0.013	0.080
	2	0.026	0.051	0.084	0.031	0.191
	3	0.064	0.125	0.208	0.076	0.473
	4	0.034	0.068	0.113	0.041	0.256
	Overall	0.134	0.265	0.439	0.162	1

The individual cells are derived through multiplying the marginal totals. From these individual cells and through using equation (5.2.11) the total sample size is estimated as 229 patients. There are anticipated to be 31.1% concordant responses (from the diagonal) and so from this the discordant sample size could be estimated as 161 patients

In Chapter 4 the methodology for a parallel group trial was extended to that for cross-over trials where the sample size per arm calculated for a parallel group study was taken as the total sample for a cross-over study. Applying the same arguments to the ordinal case, using the marginal totals as the basis for the sample size calculation and (5.2.3) the sample size is estimated to be 213 patients in total or 149 discordant patients. This approach gives a sample size around 7% lower than through using (5.2.11).

Julious and Campbell [1998] highlighted that one can simplify one's calculations by ignoring the ordinal nature of the data; dichotomising the overall responses around the direction that subjects are discordant i.e. either just -1 or 1 and then using a discordant sample size formula from Chapter 2 [Connett, Smith and McHugh, 1987]

$$n_d = \frac{\left(Z_{1-\alpha/2}(\psi+1) + 2Z_{1-\beta}\sqrt{\psi}\right)^2}{(\psi-1)^2}. \qquad (5.2.12)$$

Here, the odds-ratio, ψ, is the ratio of positive to negative responses.

Before applying (5.2.12) one must first estimate ψ. For 43% of the patients the responses on the control therapy are expected to be higher than on the investigative whilst for 26% of patients the responses are expected to be lower on the control. Thus, ψ could be estimated

as 0.60 (approximately the same as 0.56 – the treatment effect from the initial calculations) and an estimate of the discordant sample size from (5.2.12) is 164 patients. A little higher than from (5.2.11).

It is worth noting that Julious and Campbell [1998] highlighted that for many instances the discordant sample size (5.2.12) would give would be quite similar to the total sample size from (5.2.11). The reason is that by ignoring the ordinal nature of the data, there is an over-estimation of the discordant sample size. This over-estimation is of a magnitude such that the discordant sample size for a binary response approximately can equate to the total sample for an ordinal. It does not, in this instance, as the response has 4 categories with 1 predominant category.

5.3.2.3. Sensitivity Analysis about the Estimates of the Population Effects Used in the Sample Size Calculations

Similarly to parallel group data to assess the sensitivity of a trial to the estimate of the variance bootstrapping can be applied to the get an estimate of an upper percentile for the sample variance. This plausibly high value for the variance can be put into – (5.2.11) written in terms of power

$$1-\beta = \Phi\left(\sqrt{n(\log OR)^2 / \text{var}(\log(OR))} - Z_{1-\alpha/2}\right), \qquad (5.2.13)$$

to get an assessment of the sensitivity of the study.

5.3.2.4. Worked Example

In the worked example from earlier suppose that the original data, which produced a variance estimate of 7.30, had been estimated from a trial with 100 patients. Bootstrapping on the observed data produced a bootstrap 95th percentile estimate of 7.90. A plausible high estimate of the variance - 8.2% higher than used in the sample size calculation. If this value were applied to (5.2.3) then the power would be reduced 87.7%. Hence, the study seems reasonably robust to assumptions about the variance used in the calculations.

5.3.2.5. Calculations Taking Accounting of the Imprecision of the Estimates of the Population Effects Used in the Sample Size Calculations

To account for the imprecision of the sample variance in the estimate of the sample size, similarly to parallel group trials, the following result can be applied

$$1-\beta=\frac{1}{0.998}\sum_{perc=0.001}^{0.998}0.5\left[\begin{array}{l}\Phi\left(\sqrt{n}(\log(OR))^2/[\text{var}(\log(OR))]_{perc}-Z_{1-\alpha/2}\right)\\+\Phi\left(\sqrt{n}(\log(OR))^2/[\text{var}(\log(OR))]_{perc}-Z_{1-\alpha/2}\right)\end{array}\right],$$

where the percentiles for the var(log(OR)) are estimated from an empirical bootstrap distribution derived from the original data on which the variance estimate was based.

5.3.2.6. Worked Example

For the same example as described previously. To allow for the fact that the variance was estimated from a hundred subjects the sample size would need to be increased by 2 to 231 patients in total.

5.4. Non-Inferiority Trials

Although this chapter will discuss sample size calculations for non-inferiority, and later equivalence, trials, for ordinal data for such trials these calculations are not recommended. The reason for this is that although the data, as collected, are ordinal in form, in many ways for non-inferiority and equivalence trials this is the wrong scale on which to base one's inference. The rationale for this is due to the objective of the trial.

For a superiority trial the objective is to assess whether two populations differ. This assessment is done primarily done through a P-value. When analysing ordinal data there are a number of ways of determining this P-value. In this chapter the concentration has been on methodologies based upon the assumption of proportional odds. Remember here one assumes that that the odds-ratio for each cumulative 2x2 are equal across all k categories i.e. $OR_1=OR_2=OR_3=\ldots=OR_k$. In practice the individual observed ORs will deviate slightly around the overall OR. However, the overall estimate, and inference, will hold.

For non-inferiority, and equivalence, trials one wishes to determine whether two populations do not differ. This assessment is primarily done through a confidence interval where, for a non-inferiority trial, one wishes to determine whether the lower bound is greater than some pre-specified non-inferiority margin. As discussed in previous chapters, this is operationally the same as doing a one sided test. However, it is the determination and interpretation of this non-inferiority margin, which is the issue here. In previous chapters the issues with determining non-inferiority margins was highlighted and in this chapter it was also highlighted how pre-specified cut-offs could be used to determine a treatment effect for designing a superiority trial.

Extending these arguments one can determine non-inferiority limits for ordinal data. This is when the crux of the problem is encountered, however, as for a non-inferiority trial if a

cut-off is used to determine the non-inferiority limit then it is about this that interpretation would need to be made. Obviously one can assume proportional odds and that $OR_1=OR_2=OR_3=\ldots=OR_k$ however, as highlighted previously in practice individual observed ORs will deviate at different cumulative cut-offs around an overall effect. This could be sub-optimal if the observed OR around the clinically meaningful cut off is approaching the non-inferiority limit.

Both the HADS and the RSCL highlight this point. For the former a score of 0-7 would indicate that a patient is assessed as 'Normal'. Whilst for the latter a score of 0-10 would indicate that a patient is a 'Non-case'. For both scales there would be no point demonstrating non-inferiority with an overall assessment of the odds-ratio if it cannot be proven at the clinically meaningful cut-offs.

To resolve such a problem obviously one could do some form of step down procedure. First test the overall odds-ratio and if non-inferior test the odds-ratio around a cut-off for non-inferiority. However, such an approach would be driven by the least efficient comparison i.e. the one on the dichotomous cut off.

In a roundabout, way, therefore, what one is saying is that for non-inferiority trials it is operationally easier to design and analyse them as if they were binary, using the methods described in Chapter 3, about the clinically meaningful cut-offs (or several dichotomous cut-offs simultaneously as for HADS). This has an obvious adverse effect on the sample size, as highlighted in the discussions on superiority trials, however, non-inferiority trials are conservative by nature and one's approach should reflect this.

The remainder of this section will briefly describe the calculations as if the trial will be designed and analysed on the ordinal scale.

5.4.1. Parallel Group Trials

5.4.1.1. Sample Sizes that are Estimated Assuming that the Population Effects are Known

Remember the following result for non-inferiority studies from Chapter 1.

$$Var\,(S) = \frac{(d-\Delta)^2}{(Z_{1-\alpha}+Z_{1-\beta})^2}. \qquad (5.3.1)$$

Here d here is the non-inferiority limit of interest, Δ is the anticipated mean difference and var(S) is the estimated sample variance for the log-odds-ratio for an ordinal response.

An estimate of the variance for the log-odds-ratio can be made from [Whitehead, 1993]

$$Var\ (S) = \frac{6}{n\left(1 - \sum_{i=1}^{k} \bar{p}_i^3\right)}, \qquad (5.3.2)$$

where $\bar{p}_i$ is the average response each outcome category. By equating (5.3.1) with (5.3.2) one requires

$$n = \frac{6\left[Z_{1-\beta} + Z_{1-\alpha}\right]^2}{\left[1 - \sum_{i=1}^{k} \bar{p}_i^3\right](\log(OR) - d)^2}, \qquad (5.3.3)$$

where d is the non-inferiority limit, k is the number of categories and log(OR) is an estimate of the difference between treatments

5.4.1.2. Sensitivity Analysis about the Variance that is used in the Sample Size Calculations

To assess the sensitivity of the study to the variance used in the calculations (5.3.3) could be re-written in terms of power as

$$1 - \beta = \Phi\left(\sqrt{n\left[1 - \sum_{i=1}^{k} \bar{p}_i^3\right](\log OR - d)^2 / 6} - Z_{1-\alpha}\right). \qquad (5.3.4)$$

The power could then be assessed *a priori* to a high plausible value of the variance, determined through bootstrapping, to determine the studies sensitivity to the assumptions about the sample variance.

5.4.1.3. Calculations Taking Account of the Imprecision of the Variance Used in the Sample Size Calculations

To account for the imprecision in the variance estimate used in the sample size calculations the following result could be used

$$1 - \beta = \frac{1}{0.998} \sum_{perc=0.001}^{0.998} 0.5\left[\Phi\left(\sqrt{n}(\log(OR) - d)^2 / [\text{var}(\log(OR))]_{perc} - Z_{1-\alpha}\right)\right] + \Phi\left(\sqrt{n}(\log(OR) - d)^2 / [\text{var}(\log(OR))]_{perc} - Z_{1-\alpha}\right). \qquad (5.3.5)$$

The percentiles for var(logOR)) are estimated through bootstrapping.

5.4.2. Cross-over Trials

5.4.2.1. Sample Sizes that are Estimated Assuming that the Population Effects are Known

To calculate the sample size for a cross-over non-inferiority trial for ordinal data the following result could be used

$$n = \frac{\left[Z_{1-\beta} + Z_{1-\alpha}\right]^2 \text{var}(\log(OR))}{\left[\log(OR) - d\right]^2}, \qquad (5.3.6)$$

where OR and var(log(OR)) are as defined by (5.2.8) and (5.2.10) respectively.

5.4.2.2. Sensitivity Analysis About the Variance that is used in the Sample Size Calculations

To assess the sensitivity of the study to the variance used in the calculations (5.3.6) could be re-written in terms of power as

$$1-\beta = \Phi\left(\sqrt{n(\log OR - d)^2 / \text{var}(\log(OR))} - Z_{1-\alpha}\right), \qquad (5.3.7)$$

with a plausibly high value used, estimating from bootstrapping, to assess the studies sensitivity.

5.4.2.3. Calculations Taking Account of the Imprecision of the Variance Used in the Sample Size Calculations

To account for the imprecision in the variance estimate used in the sample size calculations the following result could be used

$$1-\beta = \frac{1}{0.998}\sum_{perc\ 0.001}^{0.998} 0.5\left[\begin{array}{l}\Phi\left(\sqrt{n(\log(OR) - d)^2 / [\text{var}(\log(OR))]_{perc}} - Z_{1-\alpha}\right) \\ + \Phi\left(\sqrt{n(\log(OR) - d)^2 / [\text{var}(\log(OR))]_{perc}} - Z_{1-\alpha}\right)\end{array}\right], \qquad (5.3.8)$$

where the percentiles for var(logOR)) are estimated through bootstrapping.

5.5. As Good as or Better Trials

The issues of as good as or better trials were discussed in detail in previous chapters and these arguments can be extended to ordinal data. The one issue to highlight is that in designing such trials one may be undertaking two different types of sample size calculation.

One assuming the data is binary in form for the non-inferiority calculation and one assuming it is ordinal for superiority calculations.

5.6. Equivalence Trials

The same issues for non-inferiority trials discussed in earlier this chapter generalise to sample size calculations for equivalence trials with ordinal responses. Where practical it is recommended that the data be treated as a binary response around an ordinal cut off (such as for HADS and RSCL discussed earlier) and the methodologies described in Chapter 4 for binary data be used.

5.6.1. Parallel Group Trials

5.6.1.1. Sample Sizes that are Estimated Assuming that the Population Variance is Known

5.6.1.2. General Case

Remember again that the variance about the log-odds-ratio can be defined as [Whitehead, 1993]

$$Var\,(S) = \frac{6}{n\left(1 - \sum_{i=1}^{k} \bar{p}_i^3\right)}, \qquad (5.5.1)$$

where $\bar{p}_i$ is the average response on each outcome category and k is the number of categories. Consequently an estimate of the sample size for a given power can be estimated from

$$1 - \beta = \Phi\left(\sqrt{n\left[1 - \sum_{i=1}^{k} \bar{p}_i^3\right](\log(OR) - d)^2 / 6} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{n\left[1 - \sum_{i=1}^{k} \bar{p}_i^3\right](\log(OR) + d)^2 / 6} - Z_{1-\alpha}\right) - 1, \qquad (5.5.2)$$

where d is the equivalence limit, k is the number of categories and log(OR) is an estimate of the difference between treatments

5.6.1.3. Special Case of No Treatment Difference

As with equivalence trials discussed in previous chapters when an assumption is made of no true difference between treatments the calculations are simplified - with a direct estimate of the sample size possible. With the assumption of no true difference between treatments (equivalent to OR=1) the power can be estimated from

$$1-\beta = 2\Phi\left(\sqrt{n\left[1-\sum_{i=1}^{k}\bar{p}_i^3\right]d^2/6} - Z_{1-\alpha}\right) - 1, \qquad (5.5.3)$$

whilst a direct estimate of the sample size can be obtained from

$$n = \frac{6\left[Z_{1-\beta} + Z_{1-\alpha}\right]^2}{\left[1-\sum_{i=1}^{k}\bar{p}_i^3\right]d^2}. \qquad (5.5.4)$$

5.6.1.4. Sensitivity Analysis About the Variance that is used in the Sample Size Calculations

As with superiority and non-inferiority trials discussed earlier to assess the sensitivity of a study to assumptions about the sample variance the power, for the same sample size, could be assessed from (5.5.2) using a plausibly high value of the variance. This high plausible value for the variance could be taken as a 95th percentile from a bootstrap sample.

5.6.1.5. Calculations Taking Account of the Imprecision of the Variances Used in the Sample Size Calculations

The sample size for an equivalence study accounting for the imprecision of the sample variance can be estimated from

$$1-\beta = \frac{1}{0.998}\sum_{perc=0.001}^{0.999}\frac{\eta_1+\eta_2}{2}, \qquad (5.5.5)$$

where η_A and η_B are defined as

$$\eta_1 = \Phi\left(\sqrt{n(\log(OR_{perc}) - d)^2/\left[\text{var}(\log(OR))\right]_{perc}} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{n(\log(OR_{perc}) + d)^2/\left[\text{var}(\log(OR))\right]_{perc}} - Z_{1-\alpha}\right) - 1$$

$$\eta_2 = \Phi\left(\sqrt{n(\log(OR_{perc}) - d)^2/\left[\text{var}(\log(OR))\right]_{perc+0.001}} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{n(\log(OR_{perc}) + d)^2/\left[\text{var}(\log(OR))\right]_{perc+0.001}} - Z_{1-\alpha}\right) - 1$$

As discussed previously in this chapter the percentiles for the variance used in the calculation are estimated through bootstrapping.

5.6.2. Cross-over Trials

5.6.2.1. Sample Sizes that are Estimated Assuming that the Population Variance is Known

5.6.2.2. General Case

An estimate of the sample size for a given power can be estimated from

$$1-\beta = \Phi\left(\sqrt{(\log(OR)-d)^2/[\mathrm{var}(\log(OR))]} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{(\log(OR)+d)^2/[\mathrm{var}(\log(OR))]} - Z_{1-\alpha}\right) - 1, \quad (5.5.6)$$

where OR and var(log(OR)) are as defined by (5.2.8) and (5.2.10) respectively.

5.6.2.3. Special Case of No Treatment Difference

For the special case of no true difference between treatments (equivalent to OR=1) the power can be estimated from

$$1-\beta = 2\Phi\left(\sqrt{d^2/[\mathrm{var}(\log(OR))]} - Z_{1-\alpha}\right) - 1, \quad (5.5.7)$$

whilst a direct estimate of the sample size can be obtained from

$$n = \frac{6\left[Z_{1-\beta} + Z_{1-\alpha}\right]^2}{[\mathrm{var}(\log(OR))]d^2}. \quad (5.5.8)$$

5.6.2.4. Sensitivity Analysis About the Variance that is used in the Sample Size Calculations

As described through this chapter the sensitivity of a study to its variance estimate can be determined through bootstrapping (from the original data the variance is estimated from) to calculate a high plausible value for the variance. This power of the study could then be determined through (5.5.6) to assess the sensitivity of the study.

5.6.2.5. Calculations Taking Account of the Imprecision of the Variances Used in the Sample Size Calculations

The sample size for an equivalence study accounting for the impression in the sample variance can be estimated from

$$1-\beta = \frac{1}{0.998}\sum_{perc=0.001}^{0.998}\frac{\eta_1+\eta_2}{2}, \qquad (5.5.9)$$

where η_A and η_B are defined as

$$\eta_1 = \Phi\left(\sqrt{n(\log(OR_{perc})-d)^2/[\text{var}(\log(OR))]_{perc}} - Z_{1-\alpha}\right)+\Phi\left(\sqrt{n(\log(OR_{perc})+d)^2/[\text{var}(\log(OR))]_{perc}} - Z_{1-\alpha}\right)-1$$
$$\eta_2 = \Phi\left(\sqrt{n(\log(OR_{perc})-d)^2/[\text{var}(\log OR))]_{perc+0.001}} - Z_{1-\alpha}\right)+\Phi\left(\sqrt{n(\log(OR_{perc})+d)^2/[\text{var}(\log(OR))]_{perc+0.001}} - Z_{1-\alpha}\right)-1$$

where the percentiles for [var(log(OR))] are estimated through bootstrapping.

5.7. Estimation to a Given Precision

5.7.1. Parallel Group Trials

5.7.1.1. Sample Sizes that are Estimated Assuming that the Population Variance is Known

Earlier in this chapter detailed sample size derivation for efficacy trials with ordered categorical endpoints were given [Campbell, Julious and Altman, 1995; Julious, Walker, Campbell et al, 1997; Julious, George, Machin et al, 1997]. This work can be extended to trials based on precision. For ordered categorical data the difference between two regimens may also be expressed terms of an odds-ratio (OR)

$$d = OR = \frac{p_A(1-p_B)}{p_B(1-p_A)}. \qquad (5.6.1)$$

A $(1-\alpha)100\%$ confidence interval for log(d) can be estimated using the following variance [Whitehead, 1993]

$$Var(\log(d)) = \frac{6}{n\left[1-\sum_{i=1}^{k}\bar{p}_i^3\right]}. \qquad (5.6.2)$$

Therefore, for a given half confidence interval width, w, around the odds-ratio the following condition must be met to obtain the sample size per group

$$n = \frac{6\,Z_{1-\alpha/2}^2}{(\log(1-w))^2\left[1-\sum_{i=1}^{k}\bar{p}_i^3\right]}, \qquad (5.6.3)$$

where $\bar{p}_i$ are the expected mean responses for each of the categories on the scale. In fact it is an advantage to have the variance estimated from the mean responses here. This is because for estimation trials the objective is to estimate possible differences between

treatment and as such *a priori* it is reasonable to assume that the response on each treatment to be unknown. What is more likely to be known is the anticipated mean response.

5.7.1.2. Worked Example

A pilot study is being planned to estimate the odds ratio between comparator and control regimens where the primary endpoint is an ordered categorical outcome with 4 points on the scale. The wish is to quantify the odds ratio within ±55% (i.e. w=55%). It is anticipated that the mean responses across the scale are equal i.e. that $\bar{p}_1 = \bar{p}_2 = \bar{p}_3 = \bar{p}_4 = 0.25$. Thus, the sample size required is 39 patients per arm.

5.7.1.3. Sensitivity Analysis About the Variance that is used in the Sample Size Calculations

In assessing sensitivity of a precision based trial instead of interrogating the power of the study to high plausible values for the variance one instead interrogates the loss in precision. This could be done through re-writing (5.6.3) in terms of precision (assuming w<1)

$$\log(1-w) = -\sqrt{\frac{Z^2_{1-\alpha/2}\left[\text{var}(\log(OR))\right]}{n}}. \qquad (5.6.4).$$

As with other calculations in this chapter the high plausible value for the variance is calculated through bootstrapping.

5.7.1.4. Worked Example

The estimated variance used in the earlier worked example was 6.4. Suppose that this was based on data from just 25 patients. Bootstrapping produces an estimate for the 95th percentile for the variance of 6.87, a 7.4% increase. This will equate to the precision in the point estimates reducing to 56.1%.

5.7.1.5. Calculations Taking of Account the Imprecision of the Variance Used in the Sample Size Calculations

To account for the imprecision in the variance estimate for sample size calculations one could use the following result

$$\frac{1}{0.998}\sum_{perc=0.001}^{0.998} 0.5\left[\begin{array}{l}\Phi\left(\sqrt{n}(\log(OR)-d)^2/\left[\text{var}(\log(OR))\right]_{perc}-Z_{1-\alpha/2}\right)\\ +\Phi\left(\sqrt{n}(\log(OR)-d)^2/\left[\text{var}(\log(OR))\right]_{perc}-Z_{1-\alpha/2}\right)\end{array}\right] \geq 0.50, \qquad (5.6.5)$$

where the percentiles for var(log(OR)) are estimated through bootstrapping.

5.7.1.6. Worked Example

Accounting for the fact that the original variance was estimated from 25 patients in the sample size calculation would increase the sample size to 40 patients from the 39 previously calculated.

5.7.2. Cross-Over Trials

There is a big issue in the calculation of sample sizes for precision based cross-over trials in that the results from (5.2.8) and (5.3.10) require information on individual cell counts – which *a priori* one would not be expected to know but would in fact be trying to estimate.

Earlier in this chapter it was highlighted how to estimate the sample size for other types of trials (superiority, non-inferiority and equivalence) one could use the sample size for parallel group trials and take the sample size per arm to be the total sample size for a cross-over trial. This would potentially under estimate the sample size a little (empirically between 3-7%). However, as precision based trials are quite small, in absolute terms the under-estimation would be quite small and could be circumvented by adding 2 say to the calculated sample size. It is therefore recommended to use the parallel group methodologies described in this sub-section of the chapter to estimate the total sample size for a precision based trial.

5.8. Summary of Chapter 5

This chapter demonstrates how simplifying calculations through dichotomisation for superiority trials can adversely impact on the sample size. However, for non-inferiority and equivalence trials, due to their conservative objectives, it is recommended that dichotomised calculations be used.

When assessing sensitivity it was demonstrated that assuming a chi-squared distribution for the variance provides a conservative assessment of sensitivity. However, simulations demonstrated the approximation to the chi-distribution does not hold so well for small sample sizes - particularly if one category dominates making the response practically a dichotomous response. The recommendation therefore would be to use bootstrapping developed in this chapter to assess sensitivity to the variance, as this is a generic solution that makes no assumption about the parametric form of the data.

Extending the results of Chapter 2 to use a non-central t-distribution to calculate the sample size it was shown would provide a conservative estimate of the sample size. However, this chapter also demonstrated that bootstrapping developed in this chapter provide a solution to estimate the sample size that makes no assumption about the distribution of the variance. The recommendation is to use bootstrapping to calculate the sample size to account for the imprecision in the sample variance used in the calculations.

When designing a cross-over trial it was demonstrated that although a common effect size can be shared between a parallel group trial and a cross-over trial for ordinal data as for binary data unlike for binary data one can not use the parallel group results to calculate the sample size as this would lead to a potential underestimation of the sample size. It is recommended therefore that the more complicated cross-over ordinal methodology introduced in this chapter be used for sample size calculations.

6. CHAPTER 6 - ISSUES ASSOCIATED WITH CLINICAL TRIALS

6.1. Introduction

This chapter will describe a number of applied case studies that highlight issues associated with the design of clinical trials. The areas that will be covered are:

1. Adaptive Designs
2. Heteroscadicity of trials
3. Computer intensive methods
4. Designing based on a surrogate or novel endpoint
5. Individual clinical trials in context with wide clinical plans

In each case the objective of the chapter is not to give a definitive review of the methodologies but to show how their application can assist in the design of clinical studies pertinent to the issues described in the dissertation to date. For each case study in turn an introduction and summary will be given.

6.2. Adaptive Designs

6.2.1. Introduction to Adaptive Designs

When NASA launches a rocket to Mars it does not point the rocket in the general direction of its target and launch it with their fingers crossed with a vain hope that in 2 years time it will hit the red planet. It continuously monitors the course of the rocket tinkering and modifying its route to optimise the chances of success. Analogously in clinical trials why should one set up the study and then hope that the assumptions upon which the trial was designed were correct?

Throughout this dissertation the trial design assumption most investigated has been the assumption about the trial's variability. One solution to the problem of having an uncertain estimate of the variability is to be adaptive. The advantage of being adaptive is that it allows one to alter or stop the course of a study during its actual conduct such that unexpected occurrences are not encountered for the first time when the study has stopped and the final analysis undertaken. There are three approaches that one can adopt for adaptive designs [Julious 2004b].

1. Apply a group sequential design where the sample size in each group is fixed but interim analyses are undertaken to investigate the null hypothesis with a decision made at each analysis to stop the trial for success or failure or to enrol another cohort.

2. A design is applied where at fixed interim analyses the parameters used in the estimation of the sample size are re-estimated, such as the variance for Normal data, and the sample size is adjusted accordingly. The null hypothesis is not investigated.

3. A combination of 1. and 2. where at the interim analyses both the null hypothesis is investigated and the sample size is re-estimated – conditional on whether the trial is stopped for success or failure.

The first two approaches are relatively straightforward but the third is more complex, as the sample size re-estimation depends on a decision on the null hypothesis. This section will concentrate on group sequential designs and worked example will be introduced: a recently conducted two stage group sequential drug interaction study. There is a developing literature on this topic [Julious, 2004b, 2004e; Day, 2000; Browne, 1995; Zucker and Denne, 2002; Zucker, Wittes, Schabenberger et al, 1999; Proschan, Liu and Hunsberger, 2003; Friede and Kieser, 2001; Liu, Proschan and Pledger, 2002; Gould 1992, 1995a, 1995b, 2001; Gould and Shih 1992, 1998; Birkett and Day, 1994; Kieser and Friede, 2000; Wittes and Brittain, 1990]

6.2.2. Case Study

Recall from Chapter 1 a worked example was given of a bioequivalence study which failed, primarily, it was thought, due to an observed variance over twice as high about which the study was designed. This study led to the motivation of this dissertation, in particular the *a priori* investigation of the sensitivity of a study's design to assumptions about the variance used in the sample size calculations and the allowance for the imprecision of this variance used in the sample size calculations.

Another research area that dove tailed from this one study was an investigation of appropriate adaptive methodologies for critical path studies. Critical path studies being studies upon which the start or full implementation of other studies are dependent. The case study described here followed soon after the worked example described in Chapter 1 and is described by Julious [2004b]

In this case study, *a priori* it was believed that an investigative compound might interact, from a pharmacokinetic point of view, with desipramine, leading to the plan to conduct an *in vivo* study prior to the start of phase III. For this case study the no effect criteria, used to determine if the investigative compound had any effect on the drug exposure of the probe drug desipramine ($\mu_{D,I}$) compared to desipramine alone (μ_D), was 24% on the log scale. Thus, the null and alternative hypotheses for the trial were:

H_o: $\mu_T/\mu_R \leq 0.76$ or $\mu_T/\mu_R \geq 1.30$.

H_1: $0.76 < \mu_T/\mu_R < 1.30$.

Here, the "standard" 20% bioequivalence limits were not used as *a priori* because it was believed that the wider margin of (0.76, 1.30) was sufficient to declare no effect. Likewise it was believed that only the AUC had to fall within the no effect margin for no effect to be declared. The study followed the drug regulatory guidelines for drug interaction studies [FDA, 1999; CPMP, 1997].

One issue that became apparent when designing the study was that the variability observed in the pharmacokinetics of desipramine in previous studies varied quite markedly with three studies on file having within subject coefficient of variations (CVw) for AUC of 14%, 27% and 39%. There was no apparent rational for the diverse variabilities observed and so none of them could be discounted in calculations. This led to the issue of what variance to use to estimate the sample size as the CVw of 39% would lead to a sample size estimate approaching 5 times that a CVw of 14% would require.

To overcome this particular problem a two-stage group sequential design was applied. The advantage of this approach is that group sequential methodologies allow an interim analysis to be carried out on data from one cohort of subjects - where a decision can be made whether to stop the trial for success or failure or to enrol a second cohort of subjects. To allow for the fact that an interim analysis is made the overall type I error rate of the study should be maintained at 5% by the use of appropriate statistical methods. The concept of the group sequential trial is captured in Figure 6.1.

One issue to highlight with such trials is that it is essential that the stopping rule applied at the interim analysis be pre-specified.

For the case study therefore calculations were based on two one-sided tests, a type I error rate of 5% and a no effect range of 24% i.e. (0.76, 1.30). The group sample sizes were calculated assuming a true mean ratio of unity and CVw's of 27% and 39%, which gave a sample size of 30 subjects in each cohort. To ensure 30 subjects completed the study it was planned to have 34 subjects start in each cohort.

An equal allocation, 2.5%, of the type I error was spent in each cohort using a simple Bonferoni correction. This alpha spending allocation is a little conservative but it was a pragmatic allocation given the equal cohort sample sizes. With the Bonferoni correction "adjusted" 90% confidence intervals were presented such that the overall Type I error was maintained at 5% i.e. in the calculation of the confidence intervals $\alpha/2$ and not α was used. Operationally, this is equivalent to undertaking the two one-sided test procedure but then doubling the P-values calculated.

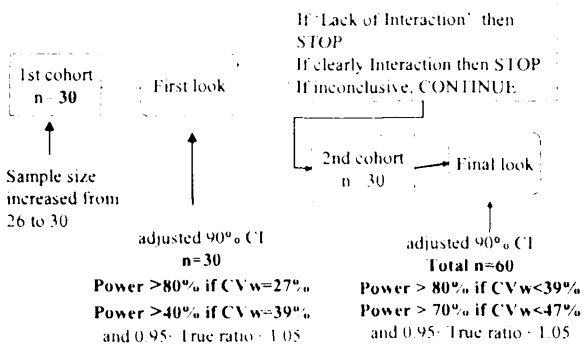
Another pragmatic decision was the choice of CVw's for sample size calculations. The pooled estimate of the CVw from the previous three studies was 28.5% however for this study it was decided just to use the two observed CVw's of 27% and 39%.

Table 6-1. Sample size and sensitivity of the sample size to assumptions about the variability and mean ratio.

CVw	Cohort	N	True Ratio		
			1.00	1.05	1.10
14%	1	30	99%	99%	99%
	2	60	99%	99%	99%
27%	1	30	90%	83%	60%
	2	60	99%	99%	90%
39%	1	30	42%	38%	28%
	2	60	90%	82%	60%

Figure 6.1 gives a description of the study design and Table 6.1 gives the breakdown of the sensitivity of the study to deviations in the assumptions about the variability and the mean difference. This table appeared in the protocol of the study. As one can see the study was quite robust to most deviations.

Figure 6-1. Concept of a group sequential trial .



The following stopping rules were applied at the interim analysis

1.	Lack of interaction: the adjusted 90% confidence interval for the ratio (S+I):S of AUCs for desipramine falls within (0.76, 1.30)
2.	Clear interaction: the ratio (S+I):S of AUCs for desipramine falls outside (0.76, 1.30)
3.	Otherwise: recruit a further 30 subjects.

A total of 34 subjects were included in the interim analysis. In the actual study at the interim analysis the adjusted 90% confidence interval for the ratio was (D+I)/D=0.94 (0.89, 1.00) with CVw=11.7% - a lower than expected variability. Therefore 'lack of interaction' of the compound on desipramine was demonstrated and the study stopped.

Had the study continued to a second cohort, the plan was to perform a "fixed sample size" analysis in each cohort separately and then to combine the two cohorts using the method described by Gould [1995b].

6.2.3. Sample Size Re-estimation - Extending the Work of the Dissertation

It is worth noting that adaptive designs and the methodologies developed in this dissertation can be married to give sample size re-estimation methodologies.

Zucker and Denne [2002] give the following sample size re-determination formula for an interim analysis

$$n = \left[\frac{2(Z_{1-\beta} + Z_{1-\alpha/2})^2 s_I^2}{d^2} \right] .IF, \qquad (6.2.1)$$

where s_I^2 is the variance estimated from the interim analysis, d is the effect size of interest, $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ are the corresponding percentiles for the standard Normal distribution for a significance level α and power $1-\beta$ and IF is the "Inflation Factor" that accounts for the imprecision in s_I^2. This correction factor is estimated through numerical methods [Zucker and Denne, 2002; Zucker, Wittes, Schabenberger et al, 1999]. However, Julious [2004e] highlighted that the inflation factor can be derived directly without the need to use numerical methods (see Chapter 2).

The direct derivation is relatively straightforward. First re-consider the "standard" way of estimating a sample size through using an equation of the form

$$n = \frac{2\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 s_I^2}{d^2}. \qquad (6.2.2)$$

To account for the fact that s_I^2 is an estimate of the population variance instead of using (6.2.2) (6.2.3) below (as given in Chapter 2), should be used to allow for the uncertainty in the sample variance used in the calculations (as assessed by its degrees of freedom)

$$n = \frac{s^2\left[TINV\left(1-\beta, \mathrm{m}, Z_{1-\alpha/2}\right)\right]^2}{d^2}, \qquad (6.2.3)$$

where *TINV(•,m,a)* is an (monotonically increasing) inverse function of the cumulative distribution function of a Students non-central t distribution with m degrees of freedom and non-centrality parameter a. Here, m is the degrees of freedom of the variance estimate, s_I^2, used in the calculations. Now the ratio of (6.2.2) and (6.2.3) depends only on α, β and m and not on s or d. Thus, by taking the ratio of (6.2.3) over (6.2.2) an inflation factor (IF) can be derived directly that adjusts the interim sample size recalculation for the uncertainty in the interim variance estimate

$$IF = \frac{\left[TINV\left(1-\beta, \mathrm{m}, Z_{1-\alpha/2}\right)\right]^2}{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}. \qquad (6.2.4)$$

Table 6.2 gives the inflation factors for different degrees of freedom calculated from (6.2.4). Equation (6.2.4) agrees with the numerical methods described by Zucker and Denne [2002] to 2 decimal places.

Table 6-2. Table of correction factors for different degrees of freedom assuming a 2 tailed type I error rate of 5% and power of 90%

Degrees of Freedom	Calculated Inflation Factors
10	1.30
15	1.19
20	1.14
25	1.11
30	1.09
40	1.07
50	1.05

Note that if one multiplies (6.2.1) by (6.2.4) one obtains (6.2.3), the direct estimate for the sample size, allowing for the degrees of freedom of the sample variance that was given in Chapter 2.

6.2.4. Summary of Adaptive Designs

Group sequential and adaptive methodologies can overcome the problems associated with imprecise estimates of variability. Although the design in the case study did not allow for sample size re-estimation the sensitivity analysis highlighted that the study was robust to deviations in the assumptions made in the sample size calculation.

Why do people not use adaptive methods more often? One reason is that it is not routine from an operational perspective to implement. In the phase I drug interaction case study described it was straightforward to implement the group sequential methodologies as the investigators conducting the trial controlled the recruitment rate. Hence, it could be fixed for subjects to arrive in a group sequential manner. For later phase trials this is not the case and the recruitment of subjects is not, usually, in the direct remit of the investigator.

There are a number of other operational issues associated with applying adaptive methodologies.

One may find that the recruitment rate may be so fast that by the time sufficient subjects have enrolled (and then had there data entered, cleaned and analysed) for interim assessment of the primary endpoint, recruitment may have completed. This point could be overcome to a degree by having a proven surrogate for the primary endpoint that would allow for adaptive decisions to be made.

Another issue is that upon starting a trial all centres do not immediately start recruiting - it may take over 6 months to initiate all centres. Thus, an early sample size review may be conducted on an unrepresentative sub sample of the current trial population, which may impact on calculations. See discussion later on heteroscadicity.

6.3. Investigating Heteroscadicity

6.3.1. Introduction to Heteroscadicity

As highlighted throughout this dissertation one of the most important components in the sample size calculation is the variance estimate used. This variance is usually estimated from retrospective data sometimes from a number of studies. To adjudicate on the relative quality of the variance, as discussed in Chapter 1, Julious [2004a] recommended considering the following aspects of the trial(s) from which the variance is obtained.

1. Design: is the study design ostensibly similar to the one you are designing? On the basic level are the data from a randomised controlled trial - observational or other data may have greater variability. If you are undertaking a multi-centre trial is the variance estimated too from a similarly designed trial? Were the endpoints similar to those you plan to use – not just the actual endpoints but were the times relative to treatment of both the outcome of interest and the baseline similar to you own?

Table 6-3. Baseline demographics and variances from 20 randomised controlled trials placebo data

Study	HAMD Entry Criteria	Number of Centres	Duration	Year	Population	Region	Phase	Sample Size	Degrees of Freedom	Variance
1	18	1	6	1984	Adult	North America	II	25	22	41.59
2	18	1	6	1985	Adult/Geriatric	North America	II	169	160	59.72
3	18	6	6	1985	Adult/Geriatric	North America	III	240	232	57.11
4	21	3	6	1986	Adult/Geriatric	North America	III	12	9	62.97
5	18	10	6	1985	Adult/Geriatric	North America	III	51	49	58.32
6	18	28	12	1991	Adult/Geriatric	North America	III	117	109	42.51
7	18	23	12	1991	Adult/Geriatric	North America	III	140	133	68.98
8	18	12	8	1992	Adult/Geriatric	North America	III	129	121	51.81
9	18	1	6	1982	Adult	Europe	III	21	19	62.44
10	15	1	6	1983	Adult/Geriatric	Europe	III	10	8	44.71
11	15	12	12	1994	Adult/Geriatric	North America	III	85	80	38.81
12	13-18	12	8	1994	Paediatric	North America	III	87	85	46.09
13	15	18	10	1994	Adult	North America	IV	43	41	60.01
14	15	20	12	1996	Adult	North America	III	101	99	61.42
15	20	20	12	1996	Adult	North America	III	110	108	61.65
16	18	29	12	1996	Geriatric	North America	III	109	105	45.54
17	20	40	8	2001	Adult/Geriatric	North America	III	146	140	58.36
18	18	1	4	1983	Adult	Europe	III	23	20	43.64
19	18	1	4	1983	Adult	Europe	III	3	1	19.32
20	18	1	4	1989	Adult	Europe	II	4	2	43.9

2. Population: is the study population similar to your own? The most obvious consideration is to ask is whether the demographics were the same but if the trial conducted was a multi centre one was it conducted in similar countries? Different countries may have different types of care (e.g. different concomitant medication) and so may have different trial populations. Was the same type of patient enrolled the same (same number of mild, moderate and severe cases)? Was it conducted covering the same seasons (relevant for conditions such as asthma)?

3. Analysis: was the same statistical analysis undertaken? This means not just the question of whether the same procedure was used for the analysis but were the same covariates fitted into the model? Was the same summary statistics used?

The quality of the variance estimate will obviously influence the sensitivity of a trial to the assumptions made about the variance and will obviously influence the strategy of an individual clinical trial. Depending on the quality of the variance estimate (or even if one has a good variance estimate) it may be advisable, as discussed earlier, in this chapter, to have some form of variance re-estimation during the trial.

Even if one has a good estimate of the variance what guarantee is there that the trial population from which the population is taken will be the same as the one the prospective trial will be drawn? One could perform two apparently identical trials (same design, same objectives, same centres) but this is not a guarantee that each trial will be drawn from the same population. For example concomitant medicines may change over time. In addition the technologies associated with the trials may change, from technologies associated with study conduct to the technology used to actually assess subjects.

The question being raised here is the heteroscadicity of trials. Empirically this can often hold for example McClung, Quessey, Julious et al [2004] observed that in COX-2 inhibitor trials in rheumatoid arthritis there was a 10% difference in placebo response for the primary endpoint of proportion of subjects being an ACR20 responder in North America (29.3%) compared to the rest of the world (40.0%). If all trials were indeed drawn from different populations then it would be problematic to base one trial design upon another. For the rheumatoid example quoted different response rates would need to be used depending on the region where the study is to be conducted.

Recommendations as to how to investigate heteroscadicity will be made through a case study.

6.3.2. Case Study

In designing a clinical trial for depression variability data were collated from a number of trials. The primary endpoint for the prospective trial was the Hamilton Depression Scale (HAMD) [Hamilton, 1960]. An appropriate estimate of the variance was thus required to use in the design of the prospective study.

6.3.2.1. The Data

The placebo data from 20 randomised controlled trials were collated for the primary endpoint of the HAMD 17 Item scale. The data sets were based on the Intent to Treat data set as this will be the primary analysis population in the future trial.

A summary of the top-level baseline demographic data for each trial is given in Table 6.2. The data span 18 years from 1983 to 2001. The studies are conducted in the two regions of Europe and North America in a number of populations. The duration of the studies varies from 4 weeks through to 12 weeks.

6.3.2.2. The Methodology

As discusses in Chapter 2 to get an overall estimate of the variance across several studies one can use the following result

$$s_p^2 = \frac{\sum_{i=1}^{k} df_i s_i^2}{\sum_{i=1}^{k} df_i}, \qquad (6.3.1)$$

where k is the number of studies, s_i^2 is the variance estimate from study i (estimated with df_i degrees of freedom) and s_p^2 is the pooled estimate of variance.

To test the heterogeneity between the study variances Bartlett's test can be applied and compared to the chi-squared distribution [Bartlett, 1937; Armitage and Berry, 1987]

$$M / C \sim \chi^2_{k-1}, \qquad (6.3.2)$$

where

$$M = \left(\sum_{i=1}^{k} df_i \right) \log(\bar{s}^2) - \sum_{i=1}^{k} df_i \log(s_i^2),$$

$$C = 1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^{k} \left(\frac{1}{v_i} \right) - \frac{1}{\sum_{i=1}^{k} v_i} \right],$$

Armitage and Berry [1987] recommended using C in the test statistic only in marginal cases as it is usually close to 1.

6.3.2.3. The Results

The pooled estimate of the variance is 55.03, which is estimated with 1543 degrees of freedom. However, there does seem to be some heterogeneity in the sample variances in the different sub-populations, given in Table 6.4, with the variability overall in the paediatric population 46.09 (on 85 degrees of freedom) and in the geriatric population 45.54 (on 105 degrees of freedom). Also, albeit on smaller populations in Europe, there seems to be a difference between the two regions of North American and Europe.

Table 6-4. Baseline demographics and variances from 20 randomised controlled trials placebo data

Population	Overall		Europe		North America	
	s_p^2	df	s_p^2	df	s_p^2	df
All	55.03	1543	50.48	50	55.19	1493
Adult	58.59	312	51.58	42	59.70	430
Adult/Geriatric	55.66	1041	44.71	8	55.74	1033
Padiatric	46.09	85	.	.	46.09	85
Geriatric	45.54	105	.	.	45.54	105

These differences, potentially, are not trivial either with differences in variances of 20% knocking on to consequent 20% difference in the sample size estimate.

Note though that this investigation of the heterogeneity ignores factors like HAMD entry criteria at baseline and study duration, which may also impact on the heterogeneity of the studies. There is no evidence of any trends by time.

For an overall test of heterogeneity, however, the Bartlett test returns a P-value of 0.561 (excluding C in the calculation) and 0.519 (including C). Thus, although the seems to be some evidence of differences in the different demographic populations the Bartlett test statistic infers that the individual studies themselves are drawn from the same population (and thus the demographic differences may be down to chance).

The data can also be examined pictorially. Data taken from a chi-squared distribution can be approximated to a Normal distribution with mean $\sqrt{2df-1}$ and variance 1. This result only technically holds for large n (and there are some small sample sizes in the case study), however, most of the studies are reasonably large. Hence, by taking away $\sqrt{2df_i-1}$ from each study (and dividing by 1) one can convert each of the variances to a scale, which approximates to the standard Normal. From these amended variances a Normal probability plot can be constructed.

Figure 6-2. Normal probability plot of the observed variances across the 20 studies in the heteroscadicity case study

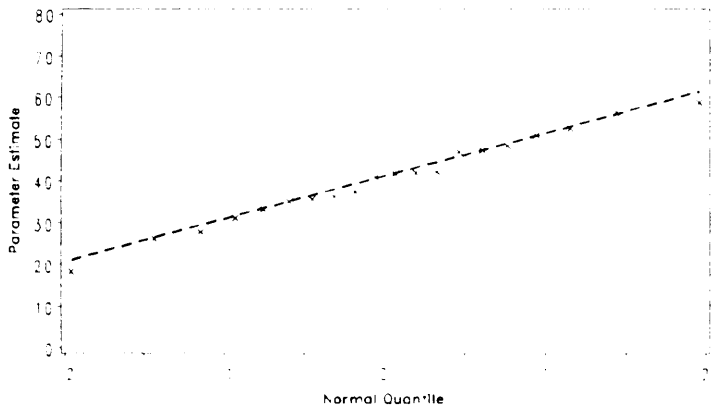


Figure 6.2 gives the Normal probability plot for the data. The bounds around the line are confidence bounds calculated using the methodologies of Friendly [1991]. Thus, this figure pictorially supports the result from the Bartlett test.

6.3.2.4. Summary of Heteroscadicity

This case study is good in that at first, with 20 studies, it seems that one has ample data upon which to estimate a variance for a sample size calculation. However, by definition the reason why there are so many studies is to interrogate different populations. Once one drills down into the data to optimise calculations for the prospective trial: same population; same study design and same region, there was fewer data to rely upon.

When assessing the data at a global level, however, there seemed to be no heteroscadicity between the studies. The evidence seems to suggest that the assumption that each study was drawn from the same population holds and that a global, pooled estimate, of the variance should be sufficient to power the prospective study.

6.4. Designing Based on a Surrogate or Novel Endpoint

6.4.1. Introduction to Designing on a Surrogate or Novel Endpoint

In the calculation of a sample size one of the most important steps is the quantification of an effect size to use in the sample size calculations. This is not straightforward for a conventional established endpoint. However, for a novel (or surrogate) endpoint it is particularly difficult as the clinical experience with using the endpoint has not been established to evaluate what a clinically meaningful difference is.

What may assist in establishing a clinically meaningful difference for a novel endpoint, however, is the association the novel endpoint may have with more established endpoints. If a clinically meaningful difference is known for these established endpoints then the association may be used to quantify an effect for the novel endpoint.

The case study below describes the calculations for quantifying an effect size in a stroke trial using a novel endpoint, the 16-question version of stroke impact scale (SIS-16) [Duncan, Lai, Bode et al, 2003; Duncan, Wallace, Lai et al, 1999]

6.4.2. Case Study

The SIS-16 assessed at three months was being considered as the primary endpoint for a stroke trial. It was felt to have advantages over conventional scales which we considered to lack sensitivity to detect possible clinical effects. For example the Barthel scale, often used as a primary endpoint, could have people with quite different disabilities getting the same score. As the SIS-16 is a comparatively novel endpoint designing a study with it as the primary endpoint led to the issue of quantifying a clinically meaningful difference. However, for other health outcomes used commonly in stroke trials, Barthel, NIHSS and Rankin, a clinically meaningful difference may be known and these have an association with SIS-16. Thus, the association between the other health outcomes and SIS-16 can be used to estimate a clinically meaningful difference for SIS-16 [Julious and Khandker, 2003].

6.4.2.1. The Methodology

The estimated treatment effect for each established health outcome was taken as an odds-ratio. A previous empirical distribution for the health outcome scores at three months was assumed to be the prospective placebo response. The health outcome score distribution on active was estimated from the placebo response under the assumption of proportional odds [Campbell, Julious and Altman, 1995; Julious, Walker, Campbell et al, 2000; Julious, George, Machin et al, 1997; Julious, George and Campbell, 1995].

To estimate an effect size for the SIS-16 following four steps were applied.

1. For each health outcome category the mean SIS-16 response was calculated.

2. Using the methods described above the expected proportions on active and placebo were estimated.

3. By multiplying the mean response for each response by the expected proportion and then summing an expected mean response was obtained for both active and placebo.

4. The difference in the mean overall responses on active and placebo were taken as an estimate of treatment effect for SIS-16 equivalent to the effect of interest for health outcomes.

Table 6-5. Worked example of the effect size estimation through associating SIS-16 with Rankin - dichotomised scale

Step 1. Calculate mean score for each Rankin category

Rankin	SIS-16 Mean
0-1	90
2-5	51

Step 2. Estimate active and placebo proportions

Rankin	SIS-16 Mean	Observed Placebo Proportion	Anticipated Active Proportion
0-1	90	0.33	0.43
2-5	51	0.67	0.57

Step 3. Multiply mean score with corresponding sample proportions and sum across categories

Rankin	SIS-16 Mean	Observed Placebo Proportion	Mean x Proportion (placebo)	Anticipated Active Proportion	Mean x Proportion (active)
0-1	90	0.33	29.7	0.43	38.7
2-5	51	0.67	34.2	0.57	29.1
Expected overall mean			63.9		67.8

Step 4. The difference taken between active and placebo means to estimate the treatment effect for SIS-16.

Rankin	SIS-16 Mean	Observed Placebo Proportion	Mean x Proportion (placebo)	Anticipated Active Proportion	Mean x Proportion (active)
0-1	90	0.33	29.7	0.43	38.7
2-5	51	0.67	34.2	0.57	29.1
Expected overall mean			63.9		67.8
Treatment Effect			67.8-63.9		=3.9

6.4.2.2. Worked Example

6.4.2.3. Dichotomised Response

For expository purposes to highlight the calculations the worked example will be first undertaken first dichotomising each of the scales around the clinically meaningful cut-offs.

Different treatment effects sizes were investigated on Barthel, Rankin and NIHSS to determine the associated effects with SIS-16:

- NIHSS - to increase the proportion of subjects classed as mild stroke (0-5%) at three months by 10%
- Rankin - to increase the proportion of subjects with Rankin score 0-1 or a Rankin score 0-2 by 10%
- Barthel - to decrease the proportion of subjects with a Barthel score 0-60 by 10%

To illustrate the calculations the calculations associating SIS-16 with increasing the proportion of subjects with Rankin score 0-1 are given in Table 6.5. The four sub-tables of Table 6.5 give the 4 steps of the calculations described in the methodology section earlier.

Table 6-6. Treatment effects for SIS-16 associated with effects on the Rankin, NIHSS and Barthel - dichotomised Scale.

	Mean Difference	Bootstrap 95% CI
Rankin 0-1	3.85	3.51 to 4.28
Rankin 0-2	4.36	3.99 to 4.74
NIHSS (0-5)	4.16	3.67 to 4.72
Barthel 0-60	6.54	4.82 to 6.25

Table 6.6 gives a summary of the effect sizes on SIS-16 associated with the three health outcomes. The confidence intervals are calculated using bootstrapping and give a measure of precision for the point estimates [Efron and Tibshirani, 1993]. Basically to do the bootstrapping repeat samples were taken with replacement from the data. For each sample an effect size was estimated and across the bootstrap sample an empirical "bootstrap" distribution was formed of effect sizes. From this bootstrap distribution the appropriate percentiles for the confidence interval were taken.

Decreasing the proportion of subjects by 10% on Barthel seems to be associated with the largest effect on SIS-16. This is probably down to the fact that in relative terms an absolute difference on this scale is quite large. It seems from these simple calculations that a mean effect of around 4 on SIS-16 would be associated with meaningful effects on the other health outcomes.

6.4.2.4. Ordinal Response

As Table 6.7 highlights by dichotomising a scale one is throwing away a fair amount of information. This is illustrated by the different mean responses in the categories collapsed into 0-1 and 2-5 in the dichotomised example earlier.

Table 6-7. Worked example of the effect size estimation through associating SIS-16 with Rankin - ordinal scale

Rankin	SIS-16 Mean	Observed Placebo Proportion	Observed Placebo Cumulative Proportion	Anticipated Active Proportion	Anticipated Active Cumulative Proportion
0	92	0.13	0.13	0.19	0.19
1	85	0.19	0.33	0.23	0.43
2	76	0.19	0.52	0.19	0.62
3	63	0.20	0.72	0.18	0.80
4	38	0.21	0.93	0.15	0.95
5	15	0.07	1.00	0.05	1.00

The same effect size is used as for the dichotomised scale but here the absolute difference is converted to an odds-ratio, which is then applied across the full scale to estimate the anticipated active response.

Table 6-8. Treatment effects for SIS-16 associated with effects on the Rankin, NIHSS and Barthel - ordinal scale.

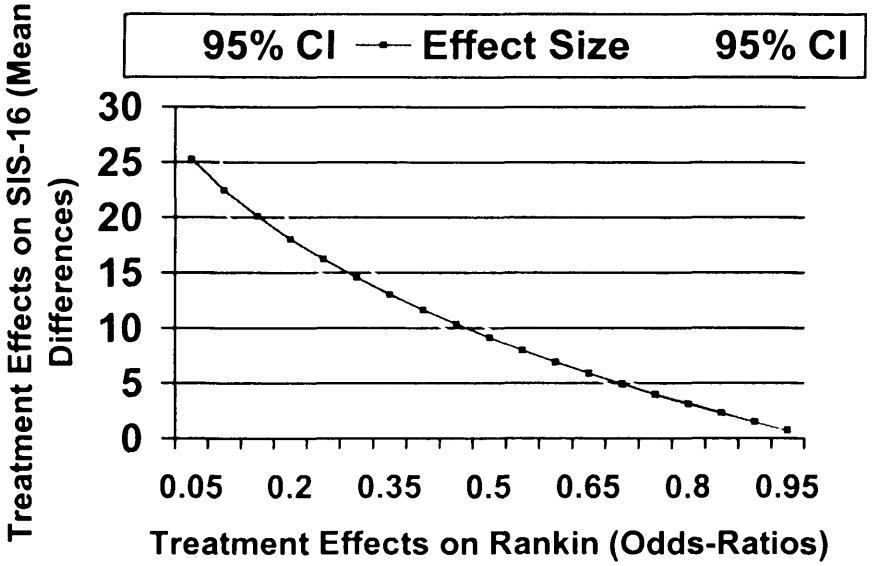
	Mean Difference	Bootstrap 95% CI
Rankin 0-1	5.20	4.78 to 5.68
Rankin 0-2	4.97	4.57 to 5.43
NIH	5.19	4.68 to 5.64
Barthel 0-60	6.63	6.04 to 7.35

Table 6.8 gives a summary of the effect sizes, and confidence intervals for these effect sizes, on SIS-16 associated with the three health outcomes. It seems that a 5-point increase in SIS-16 across the scores is associated with effects on Rankin, NIHSS and Barthel. The association with Barthel again gives the largest estimate of treatment effect.

It is reassuring that the ordinal calculations are consistent with those for the more simpler dichotomised approach. However, because of the additional information used in the calculations it is recommended that the ordinal approach be used for effect size estimation.

Although not covered in detail here the calculations can be extended for multiple calculations. In the worked example absolute differences were used to estimate odds-ratios, which were then used as a basis of effect size estimation. However, if only odd-ratios were used a variety off odds-ratios could be examined to determine different effect sizes. Figure 6.3 illustrates this point. The mid-line is the point estimate of effect size for different odds-ratios. The lower and upper lines are the bootstrap confidence intervals about this effect estimate line.

Figure 6-3. Treatment effect sizes on SIS-16 against different effects on the Rankin (odds-ratios)



6.4.3. Summary of Designing on a Surrogate or Novel Endpoint

The calculations in this case study highlight how simple calculations can enable one to estimate effect sizes for a novel endpoint if one has information on established outcomes. The same calculations could be extended to provide effect sizes for a surrogate associated with a primary outcome.

In the case study it seemed that a 5 point in improvement on SIS-16 would be associated with clinically meaningful effects on the more established outcomes.

Note though that here we are associating means with means. What one is saying is that on average if one increased the mean SIS-16 by 5 this would be associated with a mean increase in the proportion of subjects with mild stroke by 10%. No inference is being made on individual data and individual predictivity.

6.5. Computer Intensive Methods

6.5.1. Introduction to Computer Intensive Methods

Throughout this dissertation computer intensive methods have been used either to validate the sample size calculations or to undertake the sample size calculations. In Chapter 2 simulations were performed to validate the methodologies recommended; in Chapter 4 computer intensive methods were used to calculate sample sizes; while in Chapter 5 bootstrapping was applied to estimate the sample size.

In a case-by-case basis computer intensive methods can be a valuable too. If one has data from a prior, similarly designed, study then one could use these data to prospectively design future studies.

At the simplest level one could bootstrap, say, from these data to calculate the sample size. This strategy is particular useful if one is undertaking a statistical analysis for which there is no easy methodology for the calculation of the sample size. Alternatively one could use simulation-based methodologies to investigate such things as hierarchical testing strategies to minimise the Type I error.

As well as computer intensive methods to design a study one could use them to analyse the study. For example in Chapter 3 bootstrapping was recommended as way of calculation confidence intervals for the number needed to treat.

In the case study below it will be highlighted how simulations at the design stage could assist in both the design of the study and in the final statistical analysis - as well as the decision as to the most appropriate statistical analysis [Julious, 2001].

6.5.2. Case Study: Change Point Regression

When people exercise they need to produce energy and there are different metabolic pathways by which this energy is obtained (aerobic and anaerobic). For a given individual it is important to know whether a given pathway changes during exercise and, if so, when. One way of detecting this is through examining the relationship between the two metabolic variables over time while the person is exercising. In this case study a rower was to be connected to measuring equipment that reads physical responses over time. The workload was increased over time; that is, the resistance of the rowing machine to the rower was increased.

The variables considered here are those of volume of oxygen inhaled and carbon dioxide exhaled in a minute. The measurements were taken every thirty seconds up to a maximum of 17.5 minutes. What is of interest is whether there is an approximately linear relationship between the two variables or whether there is a change in slope once a critical level of oxygen inhalation is reached. The change-point represents the point at which a subject switches metabolic pathways, from aerobic to anaerobic.

The change-point regression problem was described by Quandt [1958, 1960] since when there has been an extensive literature [Shaban, 1980; Krisnaiah and Miao, 1988]. It can be applied to physiological situations where the regression slope is not expected to be constant but to change suddenly at a given point. It is this "change-point" which is of primary interest, as it may be a marker for a change in some physiological response, such as age of the menopause in a plot of bone density against age in a study of female bones [Lees, Molleson, Arnett et al, 1983], or, as here, anaerobic thresholds in subjects exercising to exhaustion [Bennett, 1988].

If the location of the change-point is known then the estimation of the parameters in the model is straightforward; however, if it is not known an extra parameter (the change-point) has to be estimated. Furthermore, the problem is no longer linear and the only way to estimate the parameters is through numerical optimisation.

This rowing model is a pharmacology model that can be used in healthy volunteers in early phase drug development, to investigate the possible pharmacological activity of a new chemical entity. More than one subject would undertake the challenge, maybe in a cross-over trial with a number of regimens or doses. New chemical entities that could be investigated in this pharmacological model are therapies that increase glycogenolysis, increasing hepatic and muscle glycogen stores, or therapies that reduce lactic acid production, such as creatinine containing products. These types of therapies would be expected to delay the change-point from aerobic to anaerobic production.

6.5.3. Location of Change-point Known

Although for the data in this planned study the location of the change-point could not be assumed to be known it is informative to discuss this special case first.

6.5.3.1. Estimation of Model

For any interval (X_0, X_1) on the real line the problem is defined as follows

$$\begin{aligned} f(x_i) &= f_1(x_i;\beta_1) \quad X_0 \le x_i \le \delta \\ &= f_2(x_i;\beta_2) \quad \delta \le x_i \le X_1 \end{aligned},$$

such that $f_1(\delta;\underline{\beta}_1) = f_2(\delta;\underline{\beta}_2)$; i.e. the slope of the relationship between y and x is constant until a point along the x-axis, δ, when it suddenly changes with no discontinuity in the regression relation. For a simple two-line linear regression this is equivalent to

$$\begin{aligned} f(x_i) \quad &= \alpha_1+\beta_1 x_i \qquad X_0 \le x_i \le \delta, \\ &= \alpha_2+\beta_2 x_i \qquad \delta \le x_i \le X_1, \end{aligned}$$

where the parameters are constrained so that $\alpha_1+\beta_1\delta=\alpha_2+\beta_2\delta$, such that the function f(x) is continuous, although not differentiable at the change-point. The least-squares estimates of the regression parameters can be derived from normal equations or alternatively the parameters for each half of the model can be estimated from [Hudson, 1966]

$$\begin{pmatrix} \dot{\underline{\beta}}_1 \\ \dot{\underline{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \overset{*}{\underline{\beta}}_1 \\ \overset{*}{\underline{\beta}}_2 \end{pmatrix} - \frac{s}{t} C^{-1} q, \qquad (6.5.1)$$

where

$$\dot{\underline{\beta}}_1' = (\underline{X}_1' \underline{X}_1)^{-1} \underline{X}_1' \underline{Y}_1,$$

$$\dot{\underline{\beta}}_2' = (\underline{X}_2' \underline{X}_2)^{-1} \underline{X}_2' \underline{Y}_2,$$

the unconstrained maximum-likelihood estimates of $\underline{\beta}_1$ and $\underline{\beta}_2$ i.e. a two-line model that is not constrained to meet at a known change-point in the range of the data. It is evident therefore that one way of deriving the least-squares estimates of the parameters is to first estimate the parameters of the two-line model where the lines are not constrained to meet at a known change-point, and then adjust these unconstrained estimates so that the two lines are constrained to meet at a known change-point.

6.5.3.2. Testing for a Regression Change when the Location of the Change-Point is Known

A two-line regression model will have residual sums of squares not larger than those for the corresponding one-line model. Therefore, to test whether the two-line model has a statistically better fit the total residual sum of squares can be used to see whether the more complicated two-line model significantly reduces the error. This leads to an F-test

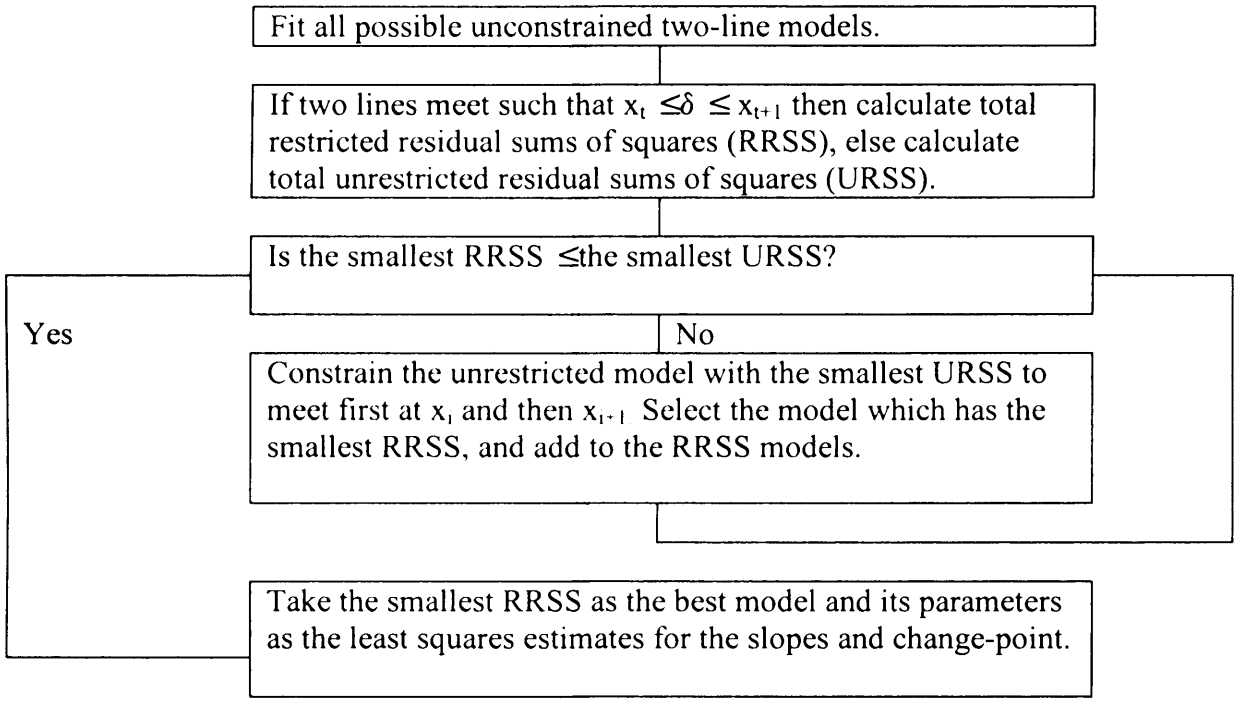
$$F = \frac{RSS_1 - RSS_2}{RSS_2 / (T - 3)}. \qquad (6.5.2)$$

Here the RSS_1 and RSS_2 are the residual sums of squares for the one- and two-line models respectively and T is the number of observations. The statistic has an F-distribution with 1 and T-3 degrees of freedom.

6.5.4. Location of Change-point Unknown

The more general case is the situation of unknown change-point which will now be described.

Figure 6-4. Algorithm to obtain an estimate of the change-point



6.5.4.1. Estimation of Model

When the location of the change-point is unknown the problem is no longer linear. The only way to estimate the parameters is through numerical optimisation and an algorithm should be used derived (Figure 6.4) to estimate all the parameters in the model [Julious, 2001].

This algorithm works as follows.

1. All unconstrained two-line models are fitted and the algorithm determines if each of these models meets within the required region on the x-axis (x_t, x_{t+1}).

2. The unconstrained models that meet within the required points are re-coded as constrained models. The algorithm then determines whether the residual error from the best fitting constrained model is smaller than the residual error from the best fitting unconstrained model. If so, then the algorithm stops and takes the constrained model with the smallest residual sums of squares as the least squares estimates. If not, then the best fitting unconstrained model is constrained, using (6.5.1) to meet at either x_t or x_{t+1} and added to the constrained models.

3. This process is repeated until one obtains the least squares estimates.

Figure 6.4 more clearly explains the iterative process.

6.5.4.2. Testing for a Regression Change when the Change-Point is Unknown - An F-statistic

An F-statistic can be derived [Worsley, 1983] that again uses the ratio of the sum of squares between the one- and two-line models,

$$F = \frac{(RSS_1 - RSS_2)/2}{RSS_2/(T-4)}. \qquad (6.5.3)$$

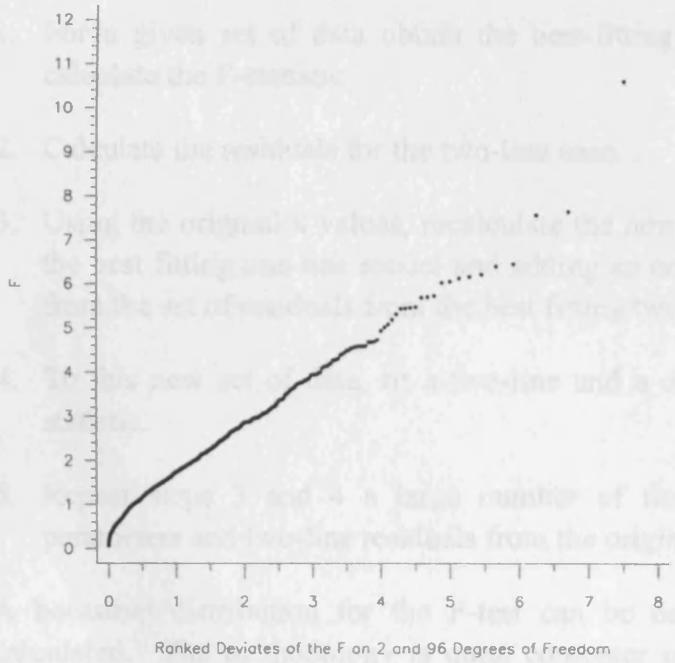
If the change-point has to be estimated, this no longer has an exact F-distribution under the null hypothesis [Hinkley, 1988], if the statistic had an exact F-distribution it would be with 2 and T-4 degrees of freedom.

6.5.4.3. Simulation Comparison of the Asymptotic F-Test

To investigate how well the distribution of the "F" statistic from (6.5.3) is approximated by the F-distribution, simulations were performed using the Interactive Matrix Language (IML) in SAS [SAS, 1985]. Simulated F values were generated by fitting a two-line model to a set of data simulated from a one-line model. The null model was assumed to have a common slope of 2 (intercept of 0), with a variance of 100. The simulation was repeated 1000 times.

If regular asymptotic theory could be applied then the "F" statistic would have an F-distribution on 2 and T-4 degrees of freedom. There were 100 points fitted by each model (x_1=1, x_2=2,...x_{100}=100), giving an F on 2 and 96 degrees of freedom. Figure 6.5 gives a probability plot of simulated F values, against ranked deviates distributed as F on 2 and 96 degrees of freedom. This plot looks fairly straight, except that there is a slight kink in the straight line at the beginning and at the end of the plot, although the slope does not seem to be unity.

Figure 6-5. Probability plot of ranked simulated F-values against ranked deviated distributed as F on 2 and 96 degrees of freedom



An F-distributed random variable with m and n degrees of freedom has expected mean and variance [Mood, Graybill and Boes, 1974]

$$mean = \frac{n}{n-2} \quad \text{for n} > 2,$$

$$\text{var}iance = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad \text{for n} > 4.$$

Therefore, for an F-test on m=2 and n=96 degrees of freedom the expected mean and variance are 1.021 and 1.088 respectively. The mean and variance of the simulated F values were 1.687 and 1.405, 65% and 30% bigger than expected. Thus, the probability plot and the deviation from the expected mean and variance suggests that asymptotic theory cannot be applied to the F-test.

6.5.4.4. Testing for a Regression Change when the Change-Point is Unknown - Bootstrapping

The F-test mentioned previously relies on assumptions regarding the distribution of the parameters. It is these assumptions, which caused the tests to fail. Efron and Gong [1983] proposed non-parametric bootstrap methods to overcome problems when using parametric tests. Bootstrap methods have been recommended for linear regression analysis [Bunke and Droge, 1984; Wu, 1986] and the extension of linear regression, change-point regression, [Hinkley, 1988; Julious, 2001] as well as a situation analogous to change-point regression, mean shift models [Hinkley and Schechtman, 1987].

The methodology in applying bootstrap methods to the change-point problem is quite straightforward:

1. For a given set of data obtain the best-fitting two-line and one-line models and calculate the F-statistic

2. Calculate the residuals for the two-line case.

3. Using the original x values, recalculate the new y values, by using the values from the best fitting one-line model and adding an error term, sampled with replacement from the set of residuals from the best fitting two-line model.

4. To this new set of data, fit a two-line and a one-line model and calculate the F-statistic.

5. Repeat steps 3 and 4 a large number of times, each time using the one-line parameters and two-line residuals from the original data.

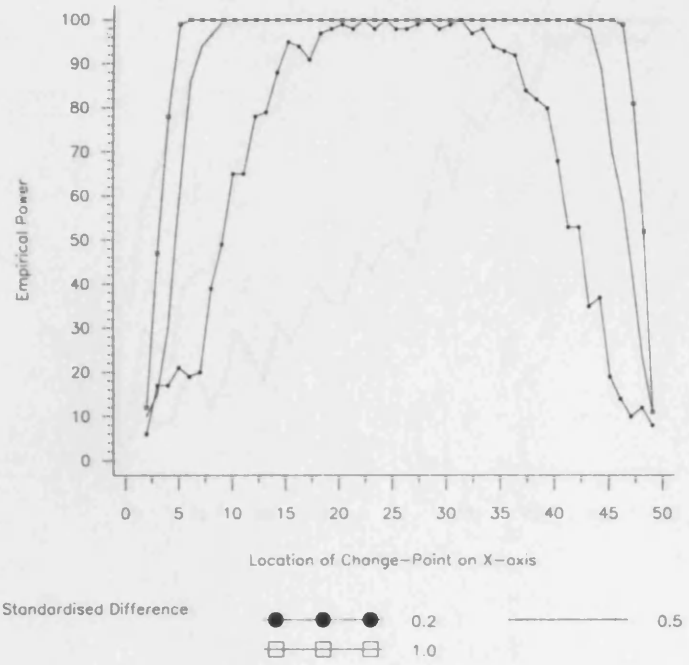
A bootstrap distribution for the F-test can be derived and a P-value can thus be calculated. The methodology is quite computer intensive although the algorithm in Figure 6.4 speeds up the estimation of the parameters.

6.5.4.5. Simulation Assessment of Bootstrapping

To investigate the properties of the bootstrap for the change-point regression problem a simulation exercise was undertaken. Simulated results were again generated using the Interactive Matrix Language (IML) in SAS (SAS Institute, 1985).

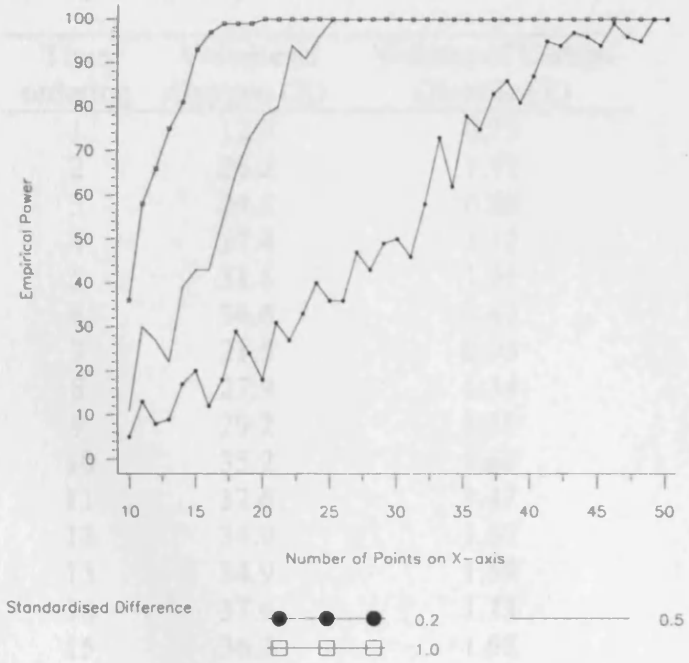
Simulations were initially undertaken to investigate the influence of the location of a change-point on the x-axis on the power of a test for various changes in slope. The data were simulated for regression changes at all the points along the x-axis from x_2 to x_{49} on a fifty-point scale. A hundred simulations were carried out for each point on the x-axis to estimate the empirical power. For each simulation a bootstrap distribution of 100 points was generated and a bootstrap significance level of 5% was chosen (thus 48x100x100 simulations were done). Figure 6.6 gives the empirical power from the simulations of regression changes at various points along the x-axis, for various values of a standardised difference d, where d is defined as $d=(\beta_1-\beta_2)/\sigma$, i.e. the difference in slopes before and after the change-point, standardised by dividing by σ, the standard deviation about the two-line model. The lines are jagged due to the noise in the simulations. The power is greatest for a change-point near the centre of the x-axis and falls towards each end of the range. The power also increases with increasing sizes of the standardised difference, d. The implication of these results is that when designing a study to investigate a possible regression change, if possible, it should be ensured that there are the same number of points before as after the change-point, to guarantee the appropriate power.

Figure 6-6. Plot of Empirical Power against Location of Regression Change for Various Slope Differences



An equivalent simulation was undertaken to assess the effect the number of points in the regression analysis had on the power. Data were simulated for regression changes at the mid-point along the x-axis for different numbers of points in the regression (from 10 to 50) for various standardised differences, d. A hundred simulations were carried out for each number of points to estimate the empirical power. For each simulation a bootstrap distribution of 100 points was generated and a bootstrap significance level of 5% was chosen (thus 40x100x100 simulations were done). Figure 6.7 gives the empirical power for a regression change for various numbers of points. From these results it seems that for a large regression change at the midpoint of the x-axis the number of points required is quite small (14 for d=1 at 80% power) with an increasing number of points required for smaller standardised differences.

Figure 6-7. Plot of empirical power against number of points in the regression change for various slope differences



6.5.4.6. Worked Example

The data for the worked example are given in Table 6.9 and are plotted in Figure 6.8. From observation of the data it seems that the variables increase over time and that there is some fluctuation due to random variation.

The best fitting single-line model with a Residual Sum Squares (RSS) of 1.072 for Carbon Dioxide Exhaled (Y_i) against Oxygen Inhaled (X_i) is

$$Y_i=-0.659+0.067X_i,$$

and the best fitting two-line model with a RSS of 0.389 is

$$Y_i=0.076+0.042X_i \qquad 12.5 \leq X_i \leq 39.46,$$

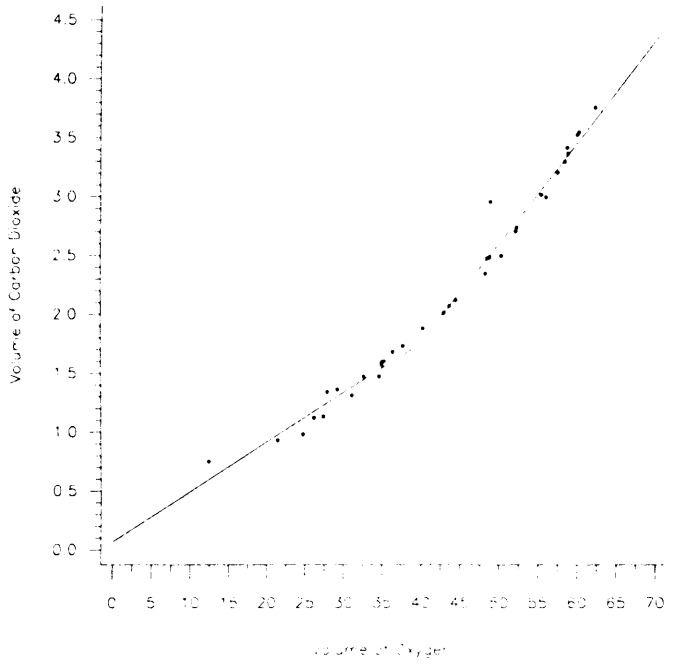
$$Y_i=-1.659+0.086X_i \qquad 39.46 \leq X_i \leq 61.8. \qquad (6.5.4)$$

The comparison of the two- and one-line models gives an F-statistic of [(1.072-0.389)/2]/[(0.389/31)]=27.21. If the statistic had an exact F-distribution it would be with 2 and 31 degrees of freedom. The bootstrap P-value (on 1000 simulations) is 0.001. The two-line model of best fit is highlighted in Figure 6.8. Visual inspection of this figure gives the impression that this model well represents the data. Although there is more than a two-fold increase in the slope between the two halves of the model, the standardised difference, at just 0.070 ((0.086-0.042)/0.624), is quite small.

Table 6-9. Data collated from measurements over time, volume of oxygen inhaled per minute (O millilitres per minute) and volume of carbon dioxide exhaled (CO2 litres per minute)

Time ordering	Volume of Oxygen (X)	Volume of Carbon Dioxide (Y)
1	12.5	0.75
2	26.2	1.12
3	24.8	0.98
4	27.4	1.13
5	31.1	1.31
6	34.6	1.47
7	21.5	0.93
8	27.9	1.34
9	29.2	1.36
10	35.2	1.60
11	32.6	1.47
12	34.9	1.57
13	34.9	1.59
14	37.6	1.73
15	36.3	1.68
16	40.1	1.88
17	42.7	2.01
18	43.4	2.07
19	44.2	2.12
20	47.9	2.35
21	49.9	2.50
22	48.1	2.48
23	48.4	2.49
24	51.7	2.71
25	51.8	2.74
26	55.5	3.00
27	54.9	3.02
28	57.0	3.21
29	57.9	3.30
30	58.3	3.37
31	58.2	3.42
32	59.5	3.53
33	59.7	3.55
34	61.8	3.76
35	48.4	2.96

Figure 6-8. Plot of volume of carbon dioxide exhaled (CO2 litres per minute) against volume of oxygen inhaled per minute (O millilitres per minute)



There is thus strong evidence to suggest that the linear relationship between the amount of carbon dioxide exhaled and oxygen inhaled changes once the amount of oxygen exceeds about 39 millilitres per minute.

6.5.5. Summary of Computer Intensive Methods

This case study highlighted the value compute intensive methods can add when designing study. They may allow one to interrogate both how to design and analyse a study.

The main theme of this dissertation is the importance of investigating *a priori* a clinical trial's robustness to the assumptions made in its design. Computing intensive methods, such as simulation, can be a valuable tool in such assessments.

6.6. Individual Trials In Context with Wider Clinical Plans

6.6.1. Introduction to Clinical Development Plans

No study is an island. In a pharmaceutical setting an individual study would form part of a clinical development plan, which will incorporate a number of studies from Phase I, pharmacology studies, through to Phase III pivotal studies. Statisticians too often concentrate on optimising individual studies but the same principles to optimise individual studies, can be extended to optimise a clinical development plan.

The following case study highlights how the application of decision sciences can optimise clinical plans [Julious and Swank, 2005]. The example given is a novel compound being developed for the acute treatment of stroke.

As with other sections in this chapter this example is not intended as a definitive review of the decision science methodologies but an indication of how decision science can assist drug development. A more detailed exposition of the theoretical work available from other work [Enas and Anderson, 2001; Burman and Senn, 2003; Senn, 1997].

6.6.2. Case Study

A project team was established charged with the development of a novel compound for the treatment of acute stroke. Developing a stroke asset is viewed as being unusually risky as

1. Only one product for acute stroke had succeeded in making it to the market to date – acteplase [NINDS rt-PA Stroke Study Group, 1995]

2. The costs of development were expected to be high relative to the commercial value.

The challenge presented to the project team was to put in place a development plan that optimised asset value whilst mitigating and controlling the risks.

6.6.2.1. Methodology

The team employed decision science techniques to create and objectively evaluate alternative different clinical development plans. The process was a team effort that required the committed involvement of representatives from marketing, project management, and clinical in addition to statistics. After alternative plans were framed, the team then evaluated the alternative plans using classic decision tree analysis [Clemen and Reilly, 2001]. The tree and the associated financial model used took into account differences in the probabilities of success for each stage of development, study costs, and launch dates.

A detailed exposition of the methodologies used in the decision analysis is beyond the scope of this dissertation. However, a brief description of the methodologies will be given.

6.6.2.2. Assessing the Value of the Asset for Different Plans

Expected net present value is the probability weighted average of the net present values (NPV) of all the possible development outcomes [Clemen and Reilly, 2001]. Net present value is the total (net) value of current and future costs and revenue expressed in

today's (present) currency. It is quite a common term used in the evaluation of assets (not just pharmaceutical) and is in fact a function in SAS.

The criteria for evaluation of the different clinical development plans are intertwined. For example, launch date can be important in its own right, but it also significantly impacts the value of a successful drug, and hence, significantly impacts eNPV.

6.6.2.3. Assessing the Probabilities of Success for Different Plans

A somewhat non-traditional approach was taken to obtain the probabilities of success assessments required for the decision tree. This approach allows for more accurate comparison of alternative development plans and makes it easier to incorporate statistical information about the ability of studies to distinguish between a successful and unsuccessful drug candidate. The process involves first assessing the team's confidence (expressed as a probability) that the drug will truly meet the safety and efficacy targets for success. This probability is the same for all the alternative development plans. Once the probability that the drug will really work has been assessed, the team then assesses the probability that each phase in the alternative clinical development plans will correctly indicate the drug works (sensitivity) or does not work (specificity).

Figure 6-9. Example of probability of success calculations

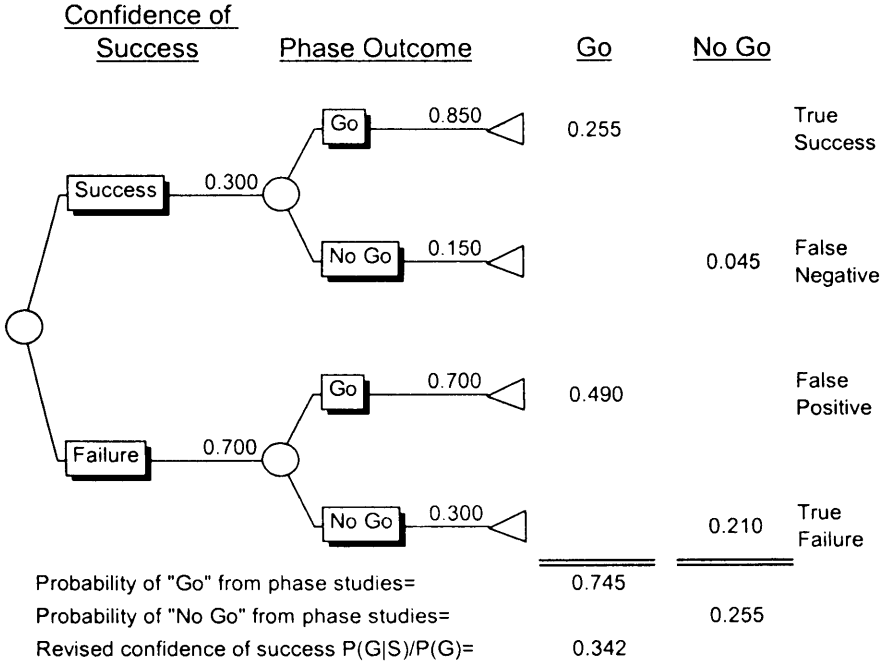


Figure 6.9 illustrates an example of the approach used to assess the probabilities of success for just one single phase's outcomes in a clinical development plan. In this example, the team's confidence the drug will really work is 0.300. The probability that the studies in this phase of development will correctly indicate "Go" to the next phase of

development when the drug really works was assessed at 0.850. The probability of a "false negative" from this phase of development is, therefore, 0.150. The probability that the studies in this phase of will correctly indicate "No Go" when the drug does not work is 0.300. The probability of a "false positive" for this phase of development is, therefore, 0.700. These probabilities might be typical of a Phase IIA or IIB study. The probability of success for each stage of development is then calculated as illustrated in the Figure 6.9. The probability of success ("Go" Outcome) is 0.745 (sum of true success and false positives).

The revised (or posterior) confidence of success is calculated through the use of Bayes' Theorem as 0.342 (0.255/0.745), a slight increase from the confidence of success prior to this study. This confidence of success would be the starting point for calculating the probability of success for the next phase of development.

6.6.3. Evaluation of the Clinical Development Plans

The team considered over 20 different alternative clinical development plans before selecting the most attractive options. So many plans were evaluated in fact that the tool the team was using had to be expanded to allow us to further evaluate alternatives. In truth though most of these plans were variations of three themes. However, there was not a clear and distinct trichotomy of plans with the team working to accentuate the strengths and to mitigate the weaknesses of individual plans. Decision Science techniques then assisted in assessing the impact of any changes.

It took a great deal of time for the team to come up with the inputs for the first few plans. However, the decision tree format for the output allowed for easy interpretation of the results. Before long the team was using the tools in real time to explore "what ifs" to see how changes in the plan would impact value, hence the number of development plans.

The focus now will be on three representative plans. Plan A is a development plan that utilises limited Phase II studies before starting two large Phase III studies.

Plan B utilises a powered imaging study in Phase II prior to the Phase III studies to reduce the risk of Phase III failure. Finally, Plan C uses an adaptive Phase II/Phase III study to reduce risk and speed development. Details of each of the plans follow. The actual numbers given in the case study have been changed to protect the innocent.

6.6.3.1. Plan A - Limited Phase II

This clinical development plan, excluding the Phase I enabling studies, consisted of two Phase II studies conducted sequentially followed by two Phase III studies conducted in parallel. The plan was of the form:

Phase IIA – A relatively small dose escalation safety study where the asset would be administered for the first time to stroke patients to establish a target dose. The team did not expect the drug to fail this study.

Phase IIB - A study primarily focussed with establishing a safety database at the chosen dose. Some preliminary efficacy data would also be obtained although the study would not be powered for this objective. Again, because of the study design and the expected "go/no go" criteria, the team expected the drug to succeed in this study.

Phase III - Two pivotal studies to assess the efficacy of the asset in a stroke population.

The rationale for this development plan was that it speeded up the entry of the asset into Phase III. This was considered desirable, as it is only in Phase III that an asset can "truly" be assessed. By design, this development plan has the lowest probability of a false negative. However, it has a high probability of an expensive Phase III failure.

Figure 6-10. Results of decision analysis for Plan A -limited Phase II

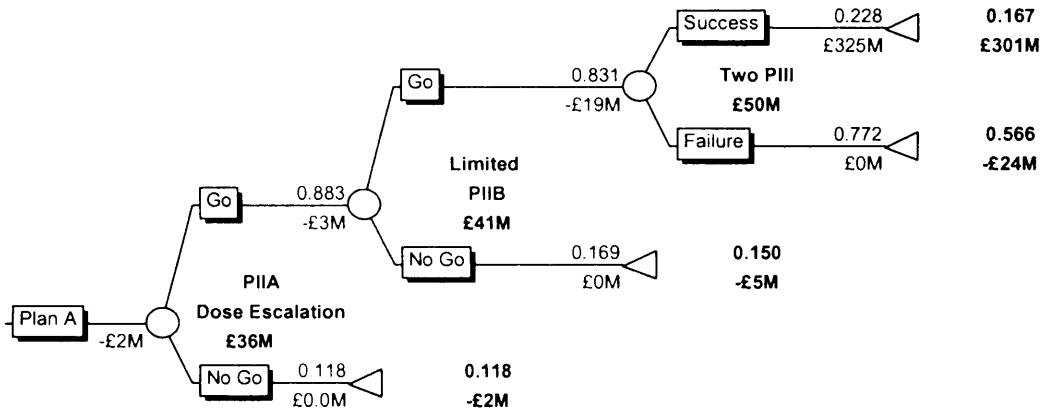


Figure 6.10 gives the decision tree analysis for Plan A. The probabilities in the figure were calculated using Bayesian methods described in section 6.6.2. The probabilities come from the team's confidence the compound will work in conjunction with the team's assessment of the sensitivities and specificities of the individual studies. The comparison of the different plans will be made in section 6.6.3.4.

6.6.3.2. Plan B - Powered Imaging Assessment in Phase II

Phase IIA – same as Plan A except more patients would be included in the IIA study to allow for better decision making prior to the start of Phase IIB. The probability of success from this study was slightly lower than the Phase IIA study in Plan A because of the higher number of patients enrolled.

Phase IIB – a study powered to assess the effect of the asset in reducing stroke infarct volume assessed through imaging. When designing the study, the team had choices about what "go/no go" criteria to use for this study. The team chose a "go/no go"

criteria that was designed to minimise the probability of a false positive (progressing a drug that will fail in Phase III). Thus, a high "Go" hurdle was set for this study. However, this had the impact of increasing the probability of a false negative (terminating a drug that works). During the actual analysis, a sensitivity analysis was done to determine the impact of different "go/no go" choices on the balance of false negative and positives from this study.

Phase III – The team postulated that the PIIB powered imaging study could be pivotal such that only one PIII pivotal trial would then be required. This possibility was allowed for in the assessment of costings etc. The sensitivity of the eNPV to this assumption was explored separately. The probability of success for Phase III was higher in this plan than in Plan A because of the risk removed by the Phase IIB imaging study.

The rationale for this development plan was that it mitigated the risk for late phase failure through assessing the asset using a surrogate in IIB – imaging. In the context of the clinical development plan this imaging study would be relatively small (requiring a fraction of the sample size of a Phase III study) although it would relatively slow in recruiting.

Figure 6-11. Results of decision analysis for Plan B - a powered imaging assessment in Phase II

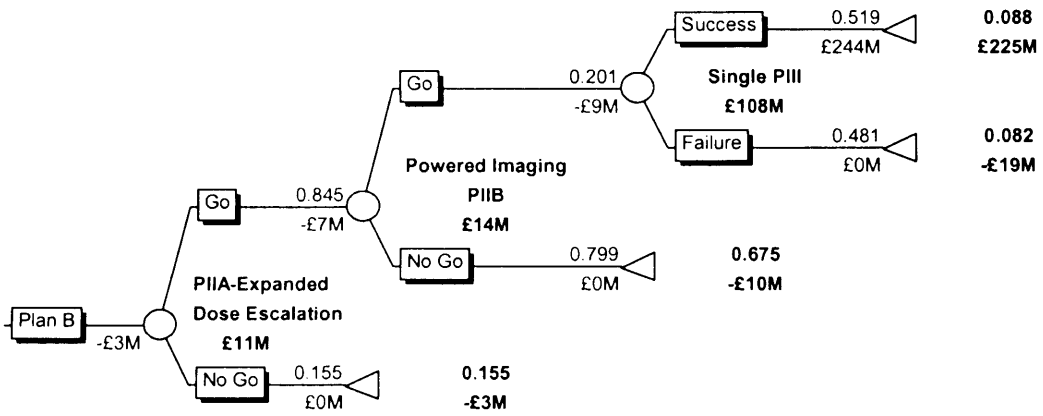


Figure 6.11 gives the decision tree analysis for Plan B. The NPV for success is much lower in Plan B because it was expected the launch date would be approximately 12 months after the launch date in Plan A.

6.6.3.3. Plan C – Adaptive Phase IIB/III Study

Phase IIB/III – An adaptive study design initially starting with two active doses (which may be altered at pre-specified interim analyses) and placebo before one dose is selected to enable the study to continue enrolment with this dose powered against placebo. The interim analysis where the dose was selected would be considered to be IIB.

Phase III - A second pivotal study would start dependent on the IIB/III study. It would start early if the interim analysis when the dose was selected was sufficiently encouraging (interim analyses to be undertaken independent of the sponsor with the sponsor remaining blind to the results). Otherwise the decision to start second Phase III would be made once the IIB/III study completes.

The rationale for this development plan was that it both allowed for a dose ranging assessment (which it was felt would reduce late phase risk) and an assessment in IIB of efficacy using the same outcomes as used in Phase III. It was felt also that the risk would be spread more evenly across Phase II and III.

Figure 6-12. Results of decision analysis for Plan C - an adaptive Phase IIb/III Study

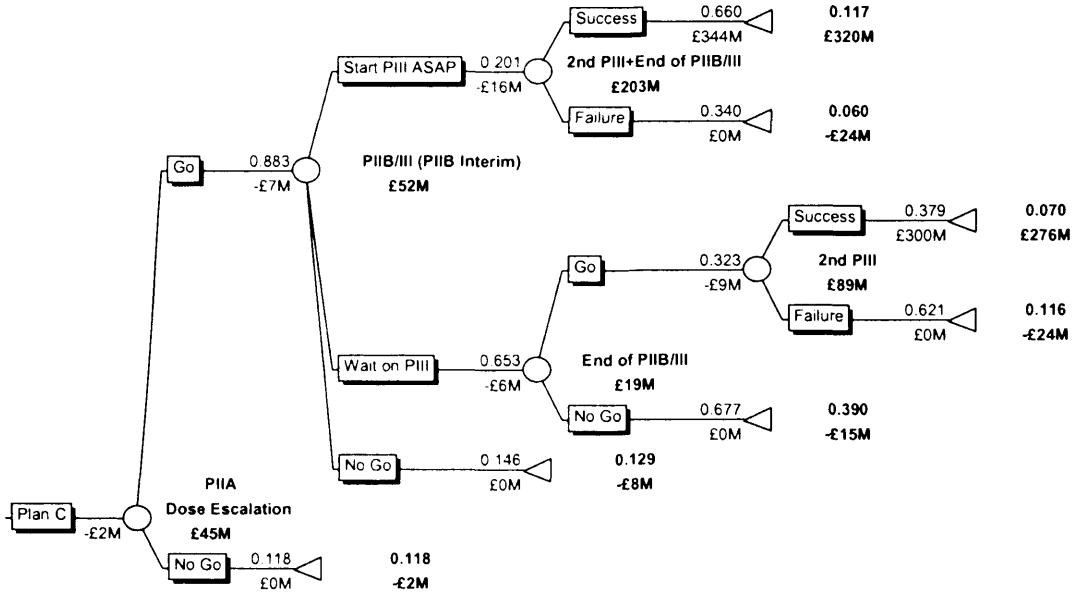


Figure 6.12 gives the decision tree analysis for Plan C. The value of success depends on the path taken to success. If the results from the PIIB/PIII are positive enough to justify starting the second PIII early, the drug can launch approximately 3 months earlier than in Plan A. However, if we must wait for the entire PIIB/PIII to complete before starting the second PIII, the launch date is approximately 4 months later than Plan A.

6.6.3.4. Results of the Evaluation

Table 6.10 and Figures 6.10-6.12 summarise the results of the decision analysis. Some figures in the table have not been previously introduced. "Probability False Negative" is simply the difference between the team's confidence (probability) the drug actually works (0.300) less the overall probability of success divided by the team's confidence

the drug actually works. "Relative Launch Month" is the expected launch date relative to Plan A. Plan C has a range of "Relative Launch Dates" depending on the likelihood the second pivotal study is started at the completion of PIIB or waits until after the end of PIIB/PIII. "eCost" is the probability weighted average (expected) development cost for each plan.

Table 6-10. Summary of clinical development plans

	Plan		
	A	B	C
	Limited PII	Phase II Imaging	Adaptive PIIB/III
Probabilities of Success			
PIIA	0.883	0.845	0.883
PIIB	0.831	0.201	0.411
PIII	0.228	0.520	0.516
Overall	0.167	0.088	0.187
Summary of Plans			
Probability PIII Failure	0.566	0.082	0.176
Probability False Negative	0.345	0.654	0.267
Relative Launch Month	0	+13	-3 to +4
NPV given success	£300M	£224M	£303M
eNPV	£36M	£11M	£45M
eCost	£19M	£10M	£16M

Based on the results of the analysis the team chose Plan C as the best plan. It has the highest expected net present value (eNPV) by over £9M, a significant amount for a project this early in development. It also has the highest overall probability of success (nearly twice that of Plan B), the potential for the earliest launch (which significantly impacts eNPV), and a reasonable expected development cost (eCost) when compared with the alternatives. Plan B would have had a lower eNPV and higher eCost if we had assume the powered PIIB imaging study could not be pivotal.

The most surprising result for the team was how poor Plan B appeared to be for this particular asset. Prior to the decision analysis, this plan had been the plan preferred by the team. Indeed Plan B may still be optimal for other assets but there were issues particular to this asset not covered in this dissertation, which impacted on Plan B (hence the high hurdle etc). However, in Plan B, it was felt by the team in this instance that there was a high probability (approximately 0.25) that a "no go" might actually have terminated a valuable therapy. In addition, the study was not a particularly "fast" nor "cheap" kill because of the difficulties associated with enrolling patients in an adequately powered study.

Before settling on its final recommended plan, the team performed a sensitivity analysis to confirm the robustness of the recommendation. The quality of the subjective probabilities assessed by the project team is always a worry. In truth there is a degree of "rubbish in, rubbish out" here. If the initial assessments are off the mark then so will be the different probabilities. However, the team felt relatively comfortable comparing the

assessments across the plans and thus, although the probability values may be off the team was confident of the relative ordering. When the team completed their sensitivity analysis, they found that Plan C was robust.

6.6.3.5. Where Should Statisticians Focus to Optimise Value

It is clear that the process for optimising clinical development plans described above offers a number of key areas where statisticians can facilitate the process and help optimise project value.

Endpoint selection: Identifying the most appropriate endpoint is critical to ensure that objectives of individual studies, and consequently the clinical plan, are met. For this case study statisticians utilised literature and existing in-house data to assist in the selection of the most appropriate endpoint(s) for the Phase III studies [Duncan, Lai, Bode et al, 2003; Khandker and Julious, 2003]. In a similar manner, they were able to provide additional insights on the use of the powered imaging study utilised in Plan B.

Study design: Individual study designs must be evaluated within the context of the overall development plan. In the case study described, the statisticians on the team introduced the concept adaptive designs to be used in Plan C to spread the risk more evenly.

Budgetary Considerations and Implications: Selection of endpoints and design impacts on sample size, which in turns impacts on budgets and timelines. Statisticians need to design all the facets of an individual development plan to assist in assessing its budgetary and time implications. In the adaptive design proposed in Plan C, the statistician also helped the team understand how many additional patients might be enrolled before the results from the interim are available to determine if an adaptive design would be feasible and practical.

Assessing Innovative Technologies: When new technologies come to the fore, the entire team needs to determine the positive impact the new technologies would have. For the case study described, Bayesian Decision Theory (not covered in this chapter) was used, through the use of priors, to help the team choose appropriate futility and early Phase III starting rules for safety and efficacy endpoints within the adaptive study.

Communication of Concepts: Some of the probabilistic concepts used in Decision Science maybe unfamiliar to certain members and would require clear explanation. This is a skill that statisticians will already have. For the case study statisticians enabled the team to determine meaningful probability assessments by illustrating the power of different studies to distinguish between successful and unsuccessful drugs.

6.6.3.6. Summary of Clinical Development Plans

In this section three different plans were discussed that were considered by the team: one which placed most of the risk in Phase III (Plan A); one which placed most of the risk in Phase II (Plan B); and one which spread the risk more evenly (Plan C). It highlighted how in the context of full clinical development plan no study is an island and how using decision sciences one can adjudicate on different options and, in this case study, give a rational for adaptive trial design methodologies over more conventional approaches.

In truth what is described in this chapter is not rocket science. However, what this decision sciences case study highlights is that "it ain't what you do it's the way that you do it" as a way to get optimal results. The process and comparisons walked through in this section were made as a consensus through optimal team working and information sharing. Through working as a team each development option was rationally assessed to come up with objective comparisons that formed the team decision.

Such decision science approaches can add considerable value to the drug development process.

7. CHAPTER 7 – SUMMARY AND CONCLUSIONS

This dissertation has reviewed the current situation with respect to sample size estimation for clinical trials. It has described the basic methodologies for sample size calculations for the most common types of trial and extended these calculations to account for the imprecision in the estimates used in the calculations.

7.1. Background

In Chapter 1 of the background to randomised clinical trials along with the basic concepts for clinical trial design were described. The most common types of clinical trials, i.e. those for superiority, equivalence, non-inferiority

7.2. Background

In Chapter 1 of the background to randomised clinical trials along with the basic concepts for clinical trial design were described. The most common types of clinical trials, i.e. those for superiority, equivalence, non-inferiority, bioequivalence and precision, were introduced and it was highlighted how the different null and alternative hypotheses impact on the calculations. The chapter then went on to explain the limitations of the conventional calculations and how, through a real world worked example, these limitations can have a severe impact on the design of a study. The concepts of sensitivity analysis and allowing for the imprecision of the estimates used in the calculations were first described.

The background to clinical trials given in this chapter was published [Julious and Zariffa, 2002] as well as the description of the different types of trial [Julious, 2004a] and the rational for trials based on precision [Julious and Patterson, 2004]. For the latter also the use of confidence intervals around individual means to interpret statistical significance between means has also been published [Julious, 2004c]. The concepts of sensitivity analysis and allowing for uncertainty of the estimates used in the trial have been published [Julious, 2004b] and presented twice at conferences at plenary invited sessions [Julious, 2001, 2002].

7.3. Normal data

In Chapter 2 standard calculations for data anticipated to take a Normal form were given. It was highlighted that the assumption that the variance used in the sample size calculation is the population variance and not a sample variance was a major assumption. The calculations for assessing the sensitivity of a trial to assumptions about the variance were given. Recommendations were also provided on obtaining the most relevant variance estimate, pertinent to the study being designed, and the

methodology was given to combine different variance estimates across several studies to get an optimal variance estimate.

A new sample size formula was given which allows for the degrees of freedom of the sample variance estimate when calculating the sample size. Initially this result was given for superiority trials. Inflation factors were provided which show the increase in sample size required for different levels of imprecision of the variance estimate (as assessed through its degrees of freedom) compared to standard calculations. If one has few degrees of freedom the inflation factors can be quite large – 30% for a study with 90% power, 5% two sided significance level and 10 degrees of freedom for the variance estimate – although falls the with greater precision. The results were compared to simulations and shown to be in agreement. With 200 degrees of freedom or more then it was shown that standard results can be used.

The chapter extended the work to other types of study. For non-inferiority and equivalence studies Bayesian methods were also introduced for the situation where as well as imprecision in the variance, imprecision in the mean difference was of interest.

For trials based around imprecision the work was compared to an alternative solution from Grieve [1991] and found to be comparable. Finally in this chapter the impact of covariates and repeat post dose measures on sample size estimation was assessed.

The standard calculations given in this chapter have appeared as a tutorial article [Julious, 2004a]. The work on investigating sensitivity have been presented at conference [Julious, 2001, 2002] and has been published [Julious, 2004b]. The solution that allows for the imprecision in the sample variance when estimating the sample size has been presented at conference [Julious, 2001, 2002] and is published [Julious and Owen, 2006]. Additionally the standard sample size calculations for precision-based trials have appeared [Julious and Patterson, 2004], as has a note on the impact of repeated post dose measures [Julious, 2000].

7.4. Binary Data

In Chapter 3 an overview of inference for binary data was given. A discussion was given as to the different types of summary statistics that could be used for binary data. It was recommended that only odds-ratios and absolute differences be used to summarise a binary response - although it was highlighted that the odds-ratio had the better mathematical properties.

For the results of Chapter 2 to be generalised to binary data the distribution for the variance around the response should follow a chi-squared distribution. It was highlighted that this was only the case for large sample sizes, which prevented the results from Chapter 2 from being extended.

Chapter 4 described the standard calculations for sample size estimation. For cross-over trials it was demonstrated that when designing a trial the parallel group sample size

methodology could be generalised to cross-over trials with the sample size per arm for a parallel group trial being used as the total sample size for a cross-over trial. It was also demonstrated that a treatment effect for a cross-over trial could be equated to a the treatment effect for a parallel group trial to assist in study design.

In addition for cross-over trials an assessment of the effect of period on sample size calculations was made. It was concluded that although possible period effects should be allowed for in any analysis their effect was small and so sample sizes could be calculated assuming there was no period effect.

For non-inferiority and equivalence studies a review of the standard calculations highlighted three quite different methods for calculations. A simulation was undertaken to compare the three approaches. From this work the it was recommended that simpler calculations [Machin, Campbell, Fayers et al, 1997] be used over the more complicated methodologies [Dunnett and Gent, 1977; Farrington and Manning, 1990; Miettinen and Nurminen, 1985; Koopman, 1984].

For binary data the factor that it was assumed to impact most on sample size calculation was a poorly estimated control response. A methodology was developed in Chapter 4 that estimated the sample size required, accounting for the imprecision in the control response, using numerical methods. The numerical methods were then extended through the use of Bayesian methodologies. It was highlighted how simple Bayesian techniques can be used to add considerable value to the calculations - particularly for non-inferiority and equivalence studies.

Finally the chapter discussed how covariates and repeated post dose measures impacted on the sample size. It was highlighted that it was a fallacy that inclusion of covariates increased sample sizes – despite increasing the variance – due to the factor their inclusion removed bias from the estimates. A bias that pulled the estimate of treatment effect towards to the null hypothesis (assuming a superiority study is being designed).

The work discussing the merits of the number needed to treat [Julious, 2002b] and the work describing calculations for non-inferiority trials [Julious, 2004d] were both given at a conference. Notes on number needed to treat and on exact confidence interval calculations are published [2005b, 2005c]

7.5. Utility of Bayesian Methods

Throughout the dissertation, but in particular in Chapter 4, a mixed approach has been undertaken whereby Bayesian methods were used in the derivation of sample sizes for clinical trials where the intention was to undertake a classical frequentist analysis. The restriction of using frequentist methods for the final analysis comes from the restriction of the work being set in a regulatory pharmaceutical setting where as discussed in Chapter 1quantification of effect is undertaken through P-values and confidence intervals.

Given that the motivation of the PhD was to account for the imprecision of previous observed estimates in the current sample size calculation an application it was logical to investigate possible Bayesian solutions. In the context of the PhD these Bayesian methods were used as a methodology as opposed to philosophy and lent in particular to a solution to the problem of sample sizes for non-inferiority trials with a binary response where the investigators beliefs as to the anticipated control response differ to what has been seen empirically. Bayesian approaches allow for the combination of these prior beliefs with the observed data to form a posterior distribution to be used in a sample size calculation that allows for the imprecision of the estimates. In turn these simple Bayesian methods can be used to assess the sensitivity of the calculations to the assumptions made both in terms of the imprecision of estimates and also to one's priors.

7.6. Ordinal Data

In Chapter 5 the sample size calculations for ordinal data were discussed. The context of the work was with respect to assessing quality of life of treatments in clinical trials. The concentration of the chapter was on trials where an assessment of treatment effect would be made through an odds-ratio through a proportional odds assumption [Whitehead, 1993].

Similarly to binary data it was demonstrated empirically that the variance around a log-odds-ratio does not follow a chi-squared distribution. Hence, the methodologies introduced in Chapter 2 could not be extended to ordinal data.

It was proposed in this chapter that bootstrapping be undertaken to both assess the sensitivity of the study to assumptions about the variance and to estimate sample size estimates accounting for imprecision in the estimate of the variance.

As with binary data it was highlighted for cross-over trials that it was possible to have estimates of treatment effect that could be thought of in terms of parallel group effects. However, unlike for binary data, the parallel group methodology could not be extended to cross-over trials as there would be an underestimation of the sample size. The more complicated results were instead recommended to be used.

Although sample size methodologies were given in this chapter for trials where the objective was to assess non-inferiority or equivalence trials it was recommended that ordinal scales be dichotomised and binary methodologies be used. The reason for this is due to the conservative nature of these trials and which hence would require a conservative hurdle to demonstrate the primary objectives.

7.7. Issues Associated with Clinical Trials

In Chapter 6 issues associated with clinical trials pertinent to the dissertation were discussed.

The dissertation has talked of issues associated with designing a trial when one has imprecise estimates and how this impacts on the sample size calculations. In Chapter 6 it was highlighted that if up front one believes the estimates to be poor then one solution could be to be adaptive in the prospectively planned trial. The adaptive design methodology was described through an applied case study.

It was also highlighted in the discussion on adaptive designs how the Normal data methodologies described in the dissertation to allow for the imprecision in the sample variance could be extended to provide sample size estimates in an interim analysis of a clinical trial.

The discussion as to the use of adaptive designs has been published [Julious, 2004b] and has been presented twice at conferences [Julious, 2001, 2002]. The result on extending the work in the dissertation to interim analyses has appeared as a note [Julious, 2004e].

Through the dissertation discussion has been made around the issues associated with designing a clinical trial based on estimates in a preceding trial. The assumption throughout has been that and study-to-study variation was purely random. However, there may be instances when the study-to-study is not random but is due to heteroscadicity of clinical trials. In Chapter 6 there was some discussion on this issue and a case study was given where possible heteroscadicity was explored. In truth the issues around heteroscadicity are complex with discussion appropriate on a case-by-case basis. This will be discussed again in the later section on areas for further work.

Chapter 6 then discussed the issue of designing a trial with an innovative or novel endpoint. The assumption in the dissertation has been that although you have imprecise estimates you at least know what clinically meaningful difference to design the study around. This chapter discussed the situation where the clinically meaningful difference is not known due to the fact a novel or surrogate endpoint is being used. The chapter described how an effect size could be estimated through its association with a previously used endpoint, which has known treatment effects. The work was illustrated through a case study from the stroke therapeutic area. This work was presented at a conference [Julious and Khandker, 2003].

Although numerical methods and simulation have been used throughout the dissertation these were used in conjunction with or to complement distributional approaches. For example in Chapter 2 simulations were performed to confirm the result accounting for the imprecision of the variance and in Chapter 4 numerical methods were used to estimate the sample size allowing for the imprecision in the control response but for each percentile used in the numerical integration the power was calculated using Normal approximation. Chapter 6 discussed how a design and analysis maybe performed if no asymptotic assumptions could be made, such that the final analysis would be performed through bootstrapping. The work was illustrated through a case study where a change-point regression was the planned analysis in a pharmacology study. This work has been published [Julious, 2001].

Finally, Chapter 6 discussed the issues where instead of considering optimising an individual study; one is considering optimising a series of studies put together into a clinical development plan. It was highlighted how simple statistical concepts such as false positive and false negatives, used regularly in individual trials, can be generalised when working to optimise a clinical plan. It was highlighted how these concepts can be framed through elementary decision science techniques. The work was highlighted through a case study for a stroke compound and is due to be published [Julious and Swank, 2005].

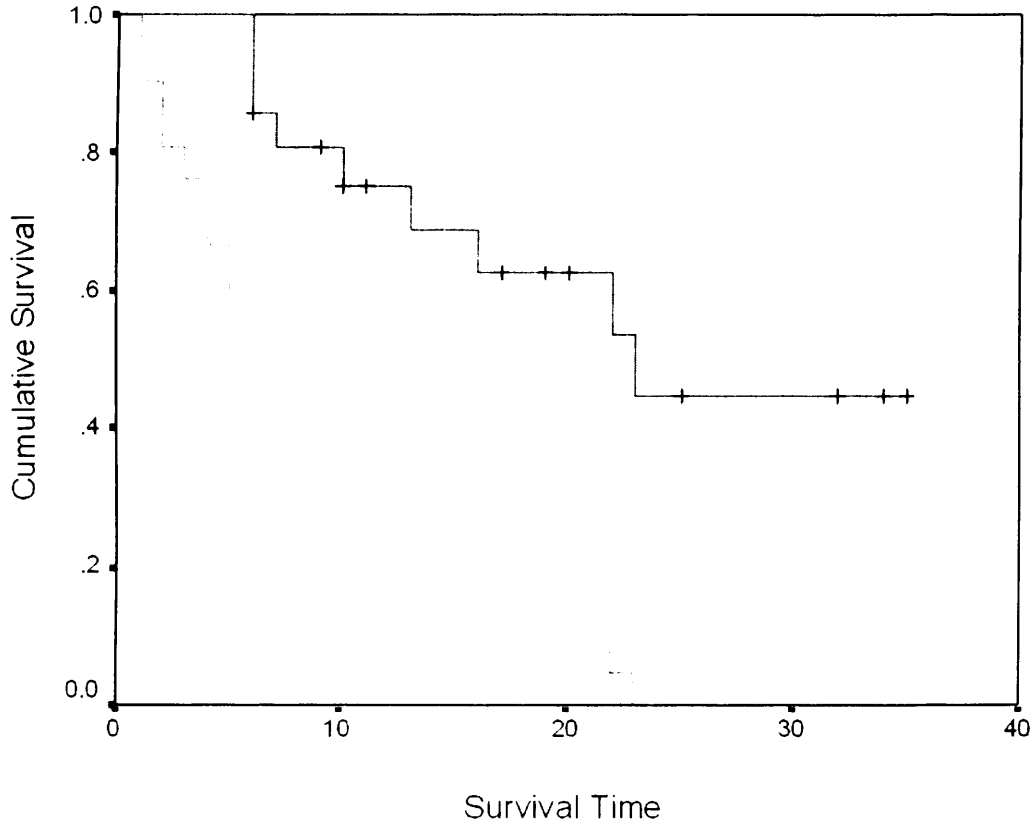
7.8. Areas for Further Work

7.8.1. Survival Data

In this dissertation methodologies have been described where the primary endpoint was anticipated to take a Normal form; be binary or be ordinal. Survival type data are common primary endpoints in trials concerned with the survival experience of the patients. Usually this survival experience is expressed in terms of survival status (e.g. alive or dead; recurred or recurrence free) and survival time (time to death; time to recurrence).

If the event of interest was observed in all subjects then the analysis, and hence design, would be relatively straightforward with a continuous primary endpoint with continuous methodologies applicable. However, studies usually finish some fixed time after start of study (e.g. 1 year) such that the event of interest is not observed in all subjects. The effect of this is that applying conventional methods for continuous endpoints would ignore subjects in whom the event was not observed.

Figure 7-1. Graphical Illustration of Survival Data.



Conversely if the data was treated as binary with the primary analysis based on a comparison of survival status by treatment time would be ignored. A survival analysis therefore accounts for the survival experience of subjects not just by investigating whether or not the event of interest has been observed in subjects but also the time to this event. Subjects where the event has not observed (the study may have stopped before the event was observed or subjects may have left before the end but not due to the defined endpoint), are treated as censored with the last observed value used in the analysis. Figure 7-1 gives a graphical illustration of survival described through a Kaplan-Meier plot [Collett, 1994].

A discussion of the generalisation of the results of the dissertation to studies where a survival endpoint feeds into the primary analysis will now be made. The emphasis will be on parallel group superiority trials although the methodologies can be extended to other designs and objectives.

A survival analysis, and hence the design of studies where there will be a survival analysis, depends on whether the event of interest is negative (e.g. death) where a proportional hazards approach will be applied or positive (e.g. cure) where an accelerated failure time approach will be applied. A series of papers by Bradburn, Clark, Love and Altman [Bradburn, Clark, Love and Altman, 2003a; Bradburn, Clark, Love and Altman, 2003b; Clark, Bradburn, Love and Altman, 2003] provide a comprehensive introduction to the analysis of time-to-event data in the context of cancer trials. They include a discussion of the relative merits of the proportional hazards and accelerated failure time model types.

7.8.1.1. Event of Primary Endpoint is Negative

Suppose the event of interest is a negative: such as death or recurrence such that the primary objective of the trial is to delay the event of happening. The primary analysis for such a response would be a log-rank test [Collett, 1994]. Now suppose the survival distributions for the two arms of the trial have instantaneous death rates of λ_A for treatment A and λ_B for treatment B. Now from this the hazard ratio (HR) is defined as

$$HR = \lambda_A/\lambda_B. \tag{7.7.1}$$

If the hazard ratio does not change with time then it can be estimated by

$$HR = \frac{\log \pi_A}{\log \pi_B}, \tag{7.7.2}$$

where π_A and π_B are two survival proportions at some fixed time point. An alternative formula for the Hazard ratio is to derive it in terms of the median survival terms for each treatment

$$HR = \frac{\log M_B}{\log M_A}, \tag{7.7.3}$$

where M_A and M_B are the median survival times on A and B respectively.

7.8.1.2. Sample Size Calculations

7.8.1.3. Method 1 – Assuming Exponential Survival

When calculating the sample sizes at the simplest level the calculations described in Chapter 4 for binary could be applied. However, this approach would ignore the survival times. A more plausible approach would be to use the methodologies in Chapter 2 for Normal data for the (probably logged) survival times. This approach would ignore the censored subjects meaning the sample size would be just for the number of events and not the total sample size. The issues allowing for the imprecision of estimates with these approaches are discussed in Chapters 2 and 4.

Alternative approaches are discussed in Machin, Campbell, Fayers and Pinol [1997]. Let T be the survival time random variable such that for treatment A one has

$$S(t) = P(T \geq t) = e^{-\lambda_A t}, \tag{7.7.4}$$

where λ_A is constant and does not change with t. From (7.7.4) one gets

$$M_A = \log_e 2/\lambda_A. \tag{7.7.5}$$

A similar result for M_B can be derived for where λ_B and hence from (7.7.3) the hazard ratio can be derived such that the number of events, E, required in each patient group is approximately

$$E = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{(\log HR)^2}. \qquad (7.7.6)$$

Note (7.7.6) involves specifying the HR only. If HR was derived from a previous trial then its imprecision could be accounted for using numerical methods described in Chapter 4. Most likely though the effect size for log(HR) would take account of other observed trials but would be based on judgment. The result (7.7.6) therefore does not use any imprecisely estimated responses.

7.8.1.4. Method 2: Proportional Hazards Only

An alternative for sample size estimation is one that assumes neither that the exponential survival or that $\lambda_A(t)$ and $\lambda_B(t)$ are constant over time, t. However, it does assume that there is a constant hazards ratio, $HR=\lambda_A(t)/\lambda_B(t)$, over time, t, such that the number of events, E, required in each patient group is approximately [Machin, Campbell, Fayers and Pinol, 1997]

$$E = \frac{(HR+1)^2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{2(HR-1)^2}. \qquad (7.7.7)$$

Note as with (7.7.6) this involves specifying HR only. As this approach makes fewer assumptions than (7.7.6) it will return slightly larger sample size estimates.

7.8.1.5. Total subjects

The results (7.7.6) and (7.7.7) give sample sizes for the number of events that are independent of the anticipated event rare in the trial. If these results were applied then the study would recruit to a specified number of events have been observed. There are obvious advantages to this approach. However, for planning purposes: for budgets; for timescales; an estimate of the total sample size would also be required.

The total sample size, n, in each group can be approximated from [Machin, Campbell, Fayers and Pinol, 1997]

$$n = \frac{2E}{2 - \pi_A - \pi_B}, \qquad (7.7.8)$$

which as well as requiring the hazard ratio specifying also requires the anticipated response rates π_A and π_B. From (7.7.2) (7.7.8) can be rewritten as

$$n = \frac{2E}{2 - e^{HR \log \pi_A} - \pi_A}. \qquad (7.7.9)$$

The result (7.7.9) is similar to the problem as described in Chapter 4 for binary data. An estimate of the anticipated response on one arm of the trial, π_A, could be available from previous studies. Using the effect size the anticipated response for the other arm, π_B, can hence be estimated. If π_A was estimated imprecisely then the sensitivity of the total sample size to this estimate could be assessed as in Chapter 2 and a sample size calculated allowing for this imprecision.

7.8.1.6. Event of Primary Endpoint is Positive

In a survival analysis sometimes the objective is to speed the event up (if the event is good). Positive events that could be investigated include: time to cure; time to remission or time to target level.

Keene [2002] describes a trial where the primary objective was time to alleviation of symptoms in an influenza trial. Figure 7.2 gives a Kaplan Meier plot from this trial. For these data the actual event rate by the end of the trial was the same for both treatments but one treatment had a faster onset of action. For trials where the objective is to speed time to event up the primary analysis would be a Generalised Wilcoxon Test [Collett, 1994].

Figure 7-2. Time to alleviation of symptoms.

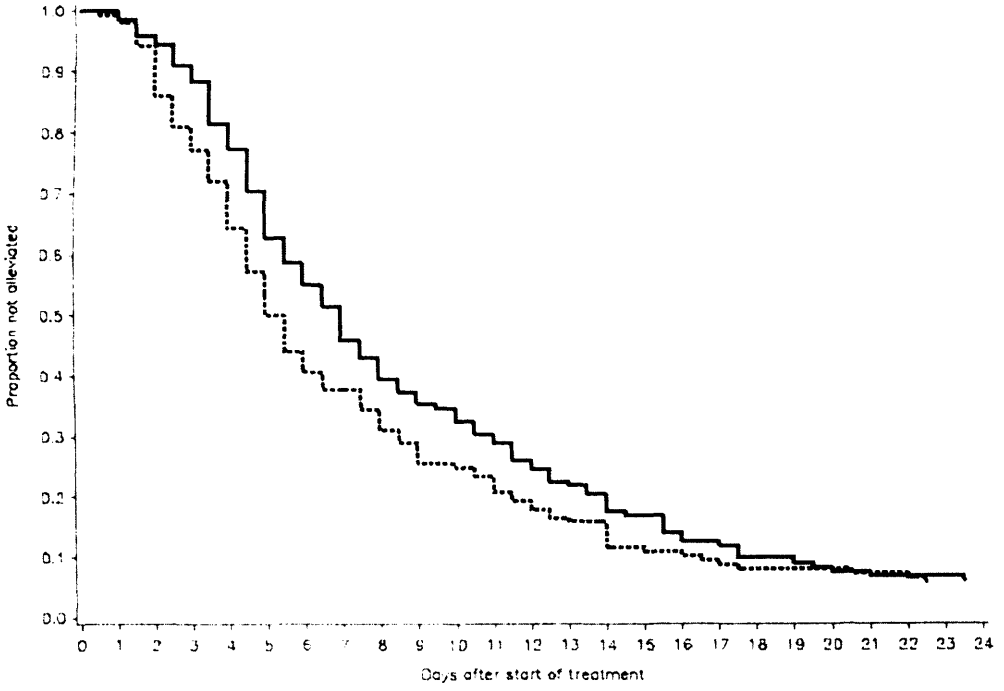


Figure 1. Kaplan-Meier plot of time to alleviation for trial 1 placebo ($n = 153$), - - - active ($n = 160$).

For the data summarised in Figure 7.2 Keene presented the results as median survival times, which were 6.0 days for placebo and 4.5 for active. These equated to a median reduction of 1.5 days with corresponding confidence intervals (calculated through bootstrapping) of 0.5 to 2.5.

An alternative approach would be to model the data through an accelerated failure time model, an approach applied by different authors for similar data [Patel, Kay and Rowell, 2006]. For data from Keene if they were expressed in terms of "acceleration factors" which are equivalent to median ratios then the estimate of effect would be 4.5/6.0=0.75. The performance of different AFT models is also evaluated in a series of examples by Kwang and Hutton [2003].

7.8.1.7. Sample Size Calculations

For data where the objective is to speed time to event there is no unique solution. If one has pre-existing data one could remove the censored subjects and applying the following methods

The method Whitehead [1993] as described in Chapter 5. The methodologies in Chapter 5 could also be applied to investigate the sensitivity of the trial to the estimates used and also to allow for the imprecision of the estimates. The disadvantage of this approach is that the calculations depend on defining an odds-ratio which is not how the data will be analysed.

The method of Noether [1987] as briefly discussed in Chapter 5. For continuous data it has the advantage of method of being distribution free, although it does not have too easily interpretable estimates of treatment effect. For discrete data, which often survival data are in trials as subjects are assessed at fixed times; there can be limitations to the method.

Probably the best method, and certainly the simplest, would be to log the survival times and assume the data take a Normal. The results from Chapter 2 could then be applied.

If there is a need to account for censoring in the sample size calculation then bootstrapping could be considered as an approach. This was discussed briefly in Chapter 6 and was used to assess sensitivity of the calculations in Chapter 5. The main advantage of this approach is that the sample size calculation may more accurately reflect the analysis. This approach should also be applied if bootstrapping is to be done in the analysis.

On a case by case level it may be optimal to use one of the methods, depending on the study, and then use another approach or two to interrogate the robustness of the sample size estimate

7.9. Cluster Randomised Trials

In the dissertation there has been a focus on pharmaceutical based regulatory trials. In such trials subjects are randomised at the individual level to receive treatment. For health technology assessments it may not always be possible to randomise at the individual level due to pragmatic considerations. Instead subjects are randomised at the level of hospital; primary care practice or practitioner level that the trial is cluster randomised.

Cluster randomised trials are therefore experiments in which intact social units rather than independent individuals are randomly allocated to intervention groups. Examples include: communities selected as the experimental unit in trials evaluating mass education programs; schools selected as the experimental unit in trials evaluating smoking prevention programs and families selected as the experimental unit in trials evaluating the efficacy of dietary interventions.

The reasons for adopting a cluster randomisation include administrative convenience; to enhance subject compliance and to avoid treatment group contamination. The latter is of particular importance as if say an education initiative was being given in a primary care setting it may not be feasible to give the intervention to one subject without subject another also being exposed. Also, if it is at the practitioner level that the intervention is being applied then it may not then be possible to individually randomise subjects and so a cluster randomisation is applied. Clustering is also an issue even in individually randomised trials where the subjects are randomised, say, to a surgical technique or to an intervention such as acupuncture. Subjects may be individually randomised but due to the finite number of practitioners there is in fact clustering. This is particularly an issue where only one arm as the cluster. Even in the pharmaceutical setting there may be a degree of clustering. If a trial had a central randomisation (i.e. not stratified by centre) but had a large number of centres relative to subjects there may be centres where subjects receive just one treatment and hence clustering.

There are disadvantages of cluster randomised trials particularly if there is a between cluster variation the presence of which as the effect of reducing the effective sample size. The extent of the problem depends on degree of within-cluster correlation and on average cluster size. There are a number of possible reasons for between-cluster variation for example: subjects frequently select the clusters to which they belong e.g. patient characteristics could be related to age or sex differences among physicians; important covariates at the cluster level affect all individuals within the cluster in the same manner e.g. differences in temperature between nurseries may be related to infection rates; individuals within clusters frequently interact and. as a result, may respond similarly e.g. education strategies or therapies provided in a group setting and finally; a tendency of infectious diseases to spread more rapidly within than among families or communities.

A consequence of the issues associated with cluster randomised trials is that standard approaches for sample size estimation and statistical analysis do not apply as standard sample size approaches would lead to an underpowered study and applying standard

statistical methods would generally tend to biased liberal -values i.e. could lead to spurious statistical significance. Cluster randomised trials also lead to the question of if you have 1000 subjects in 10 clusters for the analysis do you have a sample size of 1000 or 10? For a more complete discussion of issues associated with cluster randomised trials see Donner and Klar (2000) and Eldridge (2005).

7.9.1. Normal Data

7.9.1.1. Intra Cluster Correlation

An important consideration in designing cluster randomised clinical trials is the estimation of the intra-cluster correlation coefficient, ς. In terms of variance components the overall response variance σ^2 may be expressed as the sum of two components, i.e.

$$\sigma^2 = \sigma_B^2 + \sigma_w^2, \qquad (7.8.1)$$

where σ_B^2 is defined as the between-cluster component of variance and σ_w^2 as the within-cluster component of variance now

$$\varsigma = \frac{\sigma^2}{\sigma_B^2 + \sigma_w^2}. \qquad (7.8.2).$$

Note that $\sigma^2_W = \sigma^2(1-\varsigma)$.

7.9.1.2. Quantifying the Effect of Clustering

Consider a trial in which k clusters of size m are randomly assigned to each of an experimental and control group. Also assume the response variable Y is Normally distributed with common variance σ^2. The study is being designed as a superiority trial with the objective to test H_0:μ_A=μ_B.

Appropriate estimates of μ_A and μ_B are $\bar{x}_A^2$ and $\bar{x}_B^2$ the sample means which have the common variance

$$\frac{[1+(m-1)\varsigma]\sigma^2}{km}, \qquad (7.8.3)$$

where ς is the intra-cluster correlation coefficient (ICC), k is the number of clusters and m the sample size per cluster.

7.9.1.3. Sample Size Requirements for Cluster Randomised Designs

Suppose k clusters of size m are to be assigned to each of two intervention groups. Recall from Chapter 2 under the Normal approximation assumption the sample size for an individually randomised trial can be estimated from

$$n = \frac{2\sigma^2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{d^2}. \qquad (7.8.4)$$

To account for the effect of clustering the sample size for the number of subjects per intervention from (7.8.4) and (7.8.4) can be estimated from [Donner and Klar, 2000]

$$n = \frac{2\sigma^2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2\left[1+(m-1)\varsigma\right]}{d^2}. \qquad (7.8.5)$$

From taking the ratio of (7.8.5) over (7.8.4) an inflation factor (IF) can be estimated

$$IF=1+(m-1)\varsigma, \qquad (7.8.6).$$

to account for the cluster randomisation. Alternatively, in terms of clusters the sample size is estimate to be

$$k = \frac{2\sigma^2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2\left[1+(m-1)\varsigma\right]}{md^2} \qquad (7.8.7)$$

Actually, the results from (7.8.5) and (7.8.4) and not too dissimilar. Remember in Chapter 2 discussion was made as to the effect of covariates on the sample size such that if a single baseline was collected which was correlated with outcome by ρ the sample size could be estimated from

$$n = \frac{2\sigma^2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2\left(1-\rho^2\right)}{d^2} \qquad (7.8.8)$$

The result (7.8.8) is seldom used as for individually randomised trials a variance is used for sample size calculations from that is appropriate for the study being planned. In Chapter 1 a discussion was made as to how to assess the variance such that if the design, population and analysis from the study it is taken are similar to the one being planned then (7.8.4) could be applied. Similar principles should be considered for cluster randomised trials as if variance estimates are from trial ostensibly similar to the trial being planned then (7.8.4) and the subsequent results in Chapter 2: to assess sensitivity and to account for imprecision, can be applied.

A more detailed discussed on estimation of the intra-cluster correlation will be made in the discussion of binary data.

7.9.2. Binary Data

It is less straightforward for binary data to allow for the imprecision in the intra-cluster correlation. Recall from Chapter 4 an estimate of the sample size when one has a binary response can be estimated from

$$n = \frac{\left[\pi_A(1-\pi_A)+\pi_B(1-\pi_B)\right]\left(Z_{1-\alpha/2}+Z_{1-\beta}\right)^2}{\left(\pi_A-\pi_B\right)^2}, \qquad (7.8.9)$$

where the main determining factor in the estimate of the variance (and hence sample size) is the estimate of the control response rate π_A which, with effect size fixed, determines the anticipated response on the investigative arm π_B.

In comparison the sample size for a cluster randomised trial the sample size can be estimated from [Donner and Klar, 2000]

$$n = \frac{\left[\pi_A(1-\pi_A)+\pi_B(1-\pi_B)\right]\left(Z_{1-\alpha/2}+Z_{1-\beta}\right)^2\left[1+(m-1)\xi\right]}{\left(\pi_A-\pi_B\right)^2} \qquad (7.8.10)$$

where now on has π_A and the intra-cluster correlation ξ influencing the sample size. A number of authors have discussed the impact the intra-cluster correlation has on the design and conduct of trials [Campell, 2000; Campbell and Grimshaw, 1998; Campbell, Grimshaw and Steen, 2000]. Pertinent to the problem of allowing for the imprecision of the estimated of the intra-class in the sample size calculation is the estimation of confidence interval. Methodologies include parametric and exact methodologies [Donner and Wells, 1986; Fisher, 1925; Swiger, Harvey, Everson et al, 1964; Ukoummunne, 2003; Turner, Omar and Thompson, 2006] through to non-parametric bootstrap [Uloummunee, Davison, Gulliford et al, 2003]. While Bayesian methods have also been discussed for interval estimation [Thompson, Warn and Turner, 2004; Turner, Omar and Thompson, 2006].

The imprecision of the estimate of the intra-class correlation and its impact on sample size calculations have been discussed by a number of authors with solutions recommended from numerical approaches [Turner, Prevost Thompson, 2004] through to Bayesian [Turner, Omar and Thompson, 2001; Turner, Prevost Thompson, 2004; Spieglehalter, 2001] through to simulation [Feng and Grizzle]. While, with approaches similar to those discussed in Chapter 6, Lake, Kammann and Klar [2002] suggested adaptive sample size re-estimation approaches as a solution to imprecisely estimated intra-correlation coefficients.

In context with the dissertation the problem now in hand is to allow for imprecision in both the control response π_A and the intra-cluster correlation ξ (with fixed effect size here). The solution in the context of this dissertation is of the form

$$1-\beta = \frac{1}{10000}\sum_{i=1}^{10000}\left[\Phi\left(\sqrt{\frac{n(p_A-p_B)^2}{\left(p_{(\text{imprecise})_A}(1-p_{(\text{imprecise})_A})+p_{(\text{imprecise})_B}(1-p_{(\text{imprecise})_B})\right)\left[1+(m-1)\varsigma_{\text{imprecise}2}\right]}}-Z_{1-\alpha/2}\right)\right]_i \qquad (7.8.11)$$

where $p_{(ranperc1)_A}$ is a random percentile for the estimated control response p_A estimated using methods discussed in Chapters 3 and 4 – for example Wilson Score or Bayesian approaches. Likewise $\xi_{ranperc2}$ is a separately estimated percentile for ξ estimated using an appropriate methodology. By calculating many iterations for the percentiles for p_A and ξ and taking the average one is effectively applying a numerical percentile method; forming a distribution for the variance (accounting for the imprecision in ξ) and then numerically integrating across this to obtain the power for a given n. To obtain n for a given power one would need to iterate on n to the required power is reached. If the solutions are unstable then one would need to increase the number of permutations.

If p_A and ξ are correlated then (7.8.11) would not be appropriate. If one has actual individual data however one could bootstrap to obtain percentiles for the overall variance estimate $[\pi_A(1-\pi_A)+\pi_B(1-\pi_B)](Z_{1-\alpha/2}+Z_{1-\beta})^2[1+(m-1)\xi]$ and the methodologies in Chapter 4 could be generalised. In fact this approach could be considered anyway for all trials where one has individual data.

7.9.3. Ordinal Data

The solution for ordinal data would be similar to that for binary data although as discussed in Chapter 5 bootstrapping may be required to obtain percentiles for the variance .

7.10. Heteroscadicity of Trials

As highlighted earlier in this chapter one of the main assumptions that has been made in the development of methodologies in this dissertation is that any study-to-study variation is random and that, in terms of the variance say, any imprecision in the estimates is down purely due to insufficient degrees of freedom. However, there is reason to believe that study-to-study variability may not be random but due to trial heteroscadicity. There could be a number of possible sources of heteroscadicity such as

The technology of trials improving – the more experience gained in conducting trials in a given populations passes on learnings that enable better future trial conduct

The technology within trial improving – innovative endpoints when initially used may have greater variability than subsequent trials. An example here is imaging where but the tools of the trade and the technicians using the tools have improved – in terms of repeatability and reliability – over time

Trials being conducted in different populations – different populations could be different demographic populations or different geographic regions. Different populations may equate to heterogeneous results.

Protocol populations changing over time – the most obvious aspect here is that concomitant therapies allowed within a trial change over time as health interventions and technologies improve. Hence, a subject on placebo may be receiving different adjunct therapies in a trial designed today than 5 years a go. A consequence of this could be improved responses for subjects receiving placebo than previously would have been anticipated. This sometimes is referred to as placebo creep. There are also other factors that may also influence the heteroscadicity of trials.

An investigation as to likely heterogeneity should be undertaken prior to designing any trial, as any evidence of heteroscadicity would impact on the trial design. If there is evidence of heteroscadicity then the trial being planned would need to allow for it. Practically there would need to be an initial the sample size estimate – based on the "best guess" for the variance, say – but it would be a recommended that a sample size re-appraisal be then prospectively planned.

7.11. Contributions to Clinical Research

This dissertation has highlighted a flaw in conventional sample size calculations in that standard calculations make an assumption that the estimates uses in the calculations, such as the variance, are really known population effects. The dissertation introduced results that allowed for the imprecision in the sample estimates when estimating the sample size. For Normal data a result using the non-central t-distribution was introduced whilst for binary and ordinal data results using numerical methods were introduced.

The application of the results of the dissertation is that they provide they provide methodologies that allow for the imprecision in the estimates used in the sample size calculations. For Normal data tables of inflation factors are calculated for the main types of trial which can be used to inflate the sample size to allow for the imprecision in the variance used in the sample size calculations. For binary and ordinal data the dissertation provided solutions which are computer intensive but relatively straightforward.

When writing the sample size section for a new protocol using the methodologies from this t dissertation the assumptions made in the calculations should be clearly stated. A separate section should be included to investigate the sensitivity of the trial to these assumptions. Finally if the trial is sensitive to the assumptions actions, such as an adaptive design, should be described in the protocol to overcome these concerns.

While issues highlighted in this dissertation are particularly pertinent to innovative early phase clinical development, where, by definition, there is often little information available to design a trial the work can also be generalised to later development. For example a compound may be late in development but it may be going into a study where the primary endpoint is a novel one or alternatively but it could be going into a new patient population.

The methodologies developed in the dissertation will allow a study to be designed to allow for any possible random uncertainty in the nuisance parameters as well as allowing an investigation into any assumptions made in the design of the trial.

The work is currently being extended in particular in the area of non-inferiority trials where the trials are particular sensitive to the estimates used in calculations. The sensitivity is compounded by the fact that for active controlled non-inferiority studies the comparison to placebo is at best indirect and at worst indirect and retrospective. It is not unknown over time for placebo response to increase over time and so any indirect comparisons would need to account for this placebo creep. In this context the prior beliefs above the control response could also incorporate the beliefs of the control response over placebo to help define non-inferiority margins in the design of the trials. This is important to protect the efficacy that is observed (or has the potential to be observed) in the current active control over placebo.

The work is further being considered outside of pharmaceutical trials into health technology assessments such as in role replacement studies e.g. nurse practitioners replacing medical doctors particularly with binary responses. These are usually designed as non-inferiority trials but have the additional factor of having to allow for the imprecisely estimated intra-cluster correlation and control response rates. Bootstrapping is being considered both to combine estimates across several studies for the intra-cluster correlation and for the numerical integration when estimating the power.

A final area of extension is into there are of time to event data where the outcome is a positive event. A number of approaches are being considered for sample size calculations for example: Normal approximation; bootstrapping and Weibull each of which tie into the work undertaken in the dissertation.

8. REFERENCES

Agresti, A. (1993). Distribution free fitting of logit models with random effects for repeated categorical responses. *Statistics in Medicine* **12:**1969-1987.

Agresti, A. (1999). Modelling ordered categorical data: recent advances and future challenges. *Statistics in Medicine* **18:**2191-2207.

Agresti, A. (2003). Dealing with discreteness:making 'exact' confidence intervals for proportions, differences of proportions and odds-ratios more exact. *Statistical Methods in Medical Research* **12:**3-21.

Agresti, A. and Coull, B.A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician* **52:**119-126.

Agresti, A. and Min, Y. (2001). On sample confidence intervals for parameters in discrete distributions. *Biometrics* **57:**963-71.

Altman, D.G. (1980). Statistics and ethics in medical research III - How large a sample? *British Medical Journal* **281:**1336-8.

Altman, D.G. (1996). Better reporting of randomised trials:the CONSORT statement. *British Medical Journal* 1996 **313:**570-1.

Altman, D.G. (1998). Confidence intervals for the number needed to treat. *British Medical Journal* **317:**1309-12.

Altman, D.G. and Bland, J.M. (1999). Treatment allocation in controlled trials:why randomise? *British Medical Journal* **318:**1209.

Altman, D.G., Deeks, J.L. and Sackett D. (1998). Odds-ratios should be avoided when events are common. *British Medical Journal* **317:**1318.

Anderson, T.W. and Burnstein, H. (1967). Approximating the upper binomial confidence limit. *Journal of the American Statistical Association* **63:**857-861.

Anderson, T.W. and Burnstein, H. (1968). Approximating the lower binomial confidence limit. *Journal of the American Statistical Association* **63:**1413-1415.

Angus, J.E. and Shafer, R.E. (1934). Improved Confidence Statements for the Binomial Parameter. *The American Statistician* **38:**189-191.

Armitage, P. and Berry, G. (1987). Statistical Methods in Medical Research, 2nd Ed. Blackwell Scientific Publications, Oxford.

Bartlett, M.S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society Series A* **160:**268-282.

Bauer, P. and Kieser, M. (1996). A unifying approach for confidence intervals and testing equivalence difference. *Biometrika* **83:**934-7.

Beale, S.L. (1989). Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics* **45:**969-77.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, R., Olkin, I., et al. (1996). Improving the quality of reporting of randomised controlled trials:the CONSORT statement. *Journal of the American Medical Association* **276:**637-9

Bender, R. (2001). Calculating confidence intervals for the number needed to treat. *Controlled Clinical Trials* **22:**102-110.

Bennett, G. W. (1988). Determination of anaerobic threshold. *The Canadian Journal of Statistics* **16:**307-16.

Berger, R.L. and Hsu, J.C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* **11:**283-319.

Birkett, M.A. and Day, S.J. (1994). Internal Pilot studies for estimating sample size. *Statistics in Medicine* **13:**2455-2463.

Biswas, A. (2001). Adaptive designs for binary treatment responses in phase III clinical trials:controversies and progress. *Statistical Methods in Medical Research* **10:**353-64.

Biomarkers Definitions Working Group. (2001). Biomarkers and Surrogate Endpoints:Preferred definitions and conceptual framework. *Clinical Pharmacololy and Therapetics* **69:**89-95.

Blair, S.D., Wright, D.D.I., Blackhouse, C.M., Riddle, E. and McCullum, C.N. (1988). Sustained compression and healing of chronic venous ulcers. *British Medical Journal* **297**:1159-61.

Blyth, C.R. and Still, H.A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association* **78:**108-116.

Bradburn, M.J., Clark, T.G., Love, S.B. and Altman, D.G. (2003a). Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *British Journal of Cancer* 2003;**89**:431 – 436.

Bradburn, M.J., Clark, T.G., Love, S.B., Altman, D.G. (2003b). Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit. *British Journal of Cancer* **89**:605 – 611.

Bradford Hill A. (1990). Memories of the British streptomycin trial:the first randomized clinical trial. *Controlled Clinical Trials* **11**:77-9.

Browne R.H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Medicine* **14:**1933-1940.

Bunke, O. and Droge, B. (1984). Bootstrap and cross-validation estimates of the prediction error for linear regression models. *The Annals of Statististics* **12:**1400-24.

Bunker, J.P., Frazier, H.W. and Mosteller, F. (1994). Improving health:measuring the effects of health care. *The Millbank Quarterly* **72:**225-258.

Brush, G.G. (1988). How to choose the proper sample size. American Society for Quality Control, Milwaukee, USA.

Burman, C.F. and Senn, S. (2003). Examples of options in drug development. *Pharmaceutical Statistics* **2:**113-125.

Campbell, MJ. (2000). Cluster randomised trials in general (family) practice research. *Statistical Methods in Medical Research* **9:** 81-94

Campbell, M.K., Grimshaw, J.M. (1998). Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely ignored. *British Medical Journal* **317:** 1171-2.

Campbell, M., Grimshaw, J., Steen, N. (2000). Sample size calculations for cluster randomised trials. *Journal of Health Services Research and polity Policy* **5:**12-6

Campbell M.J., Julious S.A. and Altman D.G. (1995). Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *British Medical Journal* **311:**1145-1148.

Campbell, M.J., Julious, S.A., Walker, S.J., George, S.L. and Machin, D. (2000). A review of the use of the main quality of life measures, and sample size determination for quality of life measures, particularly in cancer trials. In:*Advanced Handbook in Evidence Based Healthcare*, Steven, A., Abrams, K.R., Brazier, J., Fitzpatrick, R. and Lilford, R.J., (eds) Sage Publications:London.

Casella, G. (1986). Refining Binomial Confidence Intervals. *The Canadian Journal of Statistics* **14:**113-129.

Chalmers, I (1998). Unbiased relevant and reliable assessments in health care. *British Medical Journal* **317:**1167-8.

Chalmers, T.C., Celano, P., Sacks, H.S. and Smith, H. (1983). Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine* **309**:1358-61.

Chan, I.S.F. (2003). Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods. *Statistical Methods in Medical Research* **12:**37-58.

Charig, C.R., Webb, D.R., Payne, S.R. and Wickham, O.E. (1986). Comparison of treatment of renal calculi by operative surgery, percutaneous nephrolithotomy and extracoporeal shock wave lithotripsy. *British Medical Journal* **292**:879-92.

Chen, J.J., Tsong, Y. and Kang, S.H. (2000). Tests for equivalence or non-inferiority between two proportions. *Drug Information Journal* **34:**569-78.

Chow, S.C., Shao, J. and Wang, H. (2002). A note on sample size calculations for mean comparisons based on noncentral t-statistics. *Journal of Pharmaceutical Statistics* **12:**441-56.

Clark, T.G., Bradburn, M.J., Love, S.B. and Altman, D.G. (2003). Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer* 2003;**89**:232 – 238.

Clayton, D and Hills, M. (1993). Statistical Models in Epidemiology. Oxford University Press, Oxford.

Clemen, R.T. and Reilly, T. (2001). Making Hard Decisions with DecisionTools, Duxbury, USA.

Clopper, C.J. and Pearson, E.S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the case of the binomial. *Biometrika* **26:**404-413.

Collett, D. (1994). Modelling survival data in medical research, Chapman and Hall, London, England

Conner, R.J. (1987). Sample size for testing differences in proportions for the paired-sample design. *Biometrics* **43:**207-211.

Connett, J.E., Smith, J.A. and McHugh, R.B. (1987). Sample size and power for pair-matched case-control studies. *Statistics in Medicine* **6:**53-59.

Cook, R.A. and Sackett, D.L. (1995). The number needed to treat:clinically useful measure of treatment effect. *British Medical Journal* **310:**452-454.

Cowell, RG, Dawid, AP, Hutchinson, TA, Roden, S and Spiegelhalter, DJ. Bayesian networks for the analysis of drug safety. The statistician 1992 42(4) 369-84

CPMP (1997). Notes for guidance on the investigation of drug interactions. Doc. CPMP/EWP/560/95. Available at URL:http://www.emea.eu.int/pdfs/human/ewp/056095en.pdf.

CPMP (1998). Notes for guidance on the investigation of bioavailability and bioequivalence. Doc. CPMP/EWP/QWP1401/98. Available at URL:http://www.emea.eu.int/pdfs/ human/ ewp/140198en.pdf.

CPMP (1999). Concept paper on the development of a committee for proprietary medicinal products (CPMP) points to consider on biostatistical methodological issues arising from recent CPMP discussions on licensing applications:choice of delta Available at URL:http://www.emea.eu.int/pdfs/human/ewp/215899en.pdf.

CPMP (2000). Points to consider on switching between superiority and non-inferiority. Doc CPMP/EWP/482/99. Available at URL:http://www.emea.eu.int/pdfs/human/ ewp/ 048299en.pdf.

CPMP (2002). Points to consider on multiplicity issues in clinical trials. Doc CPMP/EWP/908/99. Available at URL:http://www.emea.eu.int/pdfs/human/ ewp/ 090899en.pdf.

CPMP (2003). Points to consider on adjustment for baseline covariates. Doc CPMP/EWP/2863/99. Available at URL:http:// www.emea.eu.int/ pdfs/ human/ ewp/ 286399en.pdf.

CPMP (2004). Points to consider on the choice of non-inferiority margin (draft). Doc CPMP/EWP/2158/99 draft. Available at URL:http://www.emea.eu.int/ pdfs/ human/ ewp/ 215899en.pdf.

CPMP (2004). Notes for guidance on the evaluation of medicinal products indicated for the treatment of bacterial infections. Doc CPMP/EWP/558/95. Available at URL:http://www.emea.eu.int/pdfs/human/ewp/055895en.pdf.

Crow, E.L. (1956). Confidence Intervals for a Proportion. *Biometrika* **43:**423-435.

D'Agostino, R.B., Massaro, J., and Sullivan, L.M. (2003). Non-inferiority trials:design concepts and issues – the encounters of academic consultants in statistics. *Statistics in Medicine* **22:**169-86.

Daly, L. (1992). Simple SAS Macros for the Calculation of Exact Binomial and Poisson Confidence Limits. *Computational and Biological Medicine* **22:**351-361.

Dark, R., Bolland, K. and Whitehead, J. (2003). Statistical methods for ordered data based on a constrained odds model. *Biometrical Journal* **45(4):**453-70.

Davies, H.T.O., Crombie, I.K. and Tavakoli, M. (1998). When can odds-ratios mislead? *British Medical Journal* **316:**989-91.

Day, S. (1988). Clinical trial numbers and confidence intervals of pre-specified size. *The Lancet* Dec 17:1427.

Day, S. (2000). Operational difficulties with internal pilot studies to update sample size. *Drug Information Journal* **34:**461-8.

Day, S.J. and, Altman, D.G. (2000). Blinding in clinical trials and other studies. *British Medical Journal* **321**:504.

Deeks, J. (1998). Odds-ratios should be used only in case-control studies and logistic regression studies. British Medical Journal **317:**1155-6.

de Haes, J.C.J.M. and van Knippenberg, F.C.E. (1985) The quality of life of cancer patients - A review of the literature. *Social Science and Medicine* **20:**809-817.

de Haes, J.C.J.M., van Knippenberg, F.C.E. and Neijt, J.P. (1990) Measuring psychological and physical distress in cancer patients:structure and application of the Rotterdam Symptom Checklist. *British Journal of Cancer* **62:**1034-1038.

Desu, M.M. and Raghavarao, D. (1990). Sample size methodoloy. Academic Press. London.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1996). Analysis of Longitudinal Data. Oxford University Press.

Diletti, E., Hauschke, D. and Steinijans, V.W. (1991). Sample size determination for bioequivalence assessment by means of confidence intervals. *International Journal of Clinical Pharmacology, Therapy and Toxicology* **29:**1-8.

Donner, A. (1983). Approaches to sample size estimation in the design of clinical trials – a review. *Statistics in Medicine* **3:**199-214.

Donner, A. and Klar, N. (2000). Design and analysis of cluster randomization trials in health research. Arnold: London, England

Donner, A and Wells, G (1986). A comparison of confidence interval methods for the intra-class correlation coefficient. *Biometrics* **42(2)** 401-12.

Draper, N.R., and Smith, H. (1981). *Applied linear regression*, 2nd edn. Chichester: Wiley.

Duncan, P.W., Lai, S.M., Bode, R.K., Perera, S., DeRosa, J., and the GAIN Americas Investigators (2003). Stroke Impact Scale-16 A brief assessment of physical function, Neurology **60:**291-296.

Duncan, P.W., Wallace, D., Lai, S.M., Johnson, D., Emretson, S., and Jacobs, L. (1999). The Stroke Impact Scale Version 2.0 Evaluation of reliability, validity, and sensitivity to change. *Stroke* **30**:1840-3.

Dunnett, C.W. and Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometrics* **33:**593-602.

Ederer, F. and Mantel, N. (1974). Confidence limits on the ratio of two poisson variables. *American Journal of Epidemiology* **100(3):**165-7.

Edwardes, M.D. (1998). The evaluation of confidence sets with application to binomial intervals. *Statistica Sinica* **8:**393-409.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jacknife and the cross-validation. *American. Statistian*, **37:**36-48.

Efron, B. and Tibshirani, R.J. (1993). An introduction to the bootstrap. Chapman and Hall:New York.

Eldridge. S.M. (2005). Assessing, understanding an improving the efficiency of cluster randomized trials in primary care. Unpublished PhD dissertation, University of London, England.

Ellenberg, S.S. and Temple, R. (2000). Placebo-controlled trials and active control trials in the evaluation of new treatments. Part 2:Practical issues and specific cases. *Annals of Internal Medicine* **133:**464-70

Ellison, B.E. (1964). Two theorems for inferences about the Normal distribution with applications in acceptance sampling. *Journal of the American Statistical Association* **59:**89-95.

Enas, G. Andersen, J.S. (2001). Enhancing the value delivered by the statistician throughout drug discovery and development: putting statistical science into regulated pharmaceutical innovation. *Statistics in Medicine* **20:**2697-2708

Escobar, M.D. (1994). Estimating means with Dirichlet process priors. *Journal of the American Statistical Association* **89:**268-77.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90:**570-88.

Farrington, C.P. and Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine* **9:**1447-54.

Fayers, P., Ashby, D. and Parmar, M. (1997). Tutorial in Biostatistics: Bayesian Data Monitoring in Clinical Trials. *Statistics in Medicine* **16:** 1413-30.

Fayers, P. and Machin, D. (1995). Sample size:how many patients are necessary? *British Journal of Cancer* **72:**1-9.

Fayers, P.M. and Machin, D. (2000). *Quality of Life:Assessment, Analysis and Interpretation.* John Wiley:Chichester.

FDA (1992). Points to consider. Clinical evaluation of anti-infective drug products. Available at URL:http://www.fda.gov/cder/guidance/old043fn.pdf.

FDA (1997). Draft guidance for industry. Food-effect bioavailability and bioequivalence studies. Available at URL :http://www.fda.gov /cder /guidance /1719dft.pdf.

FDA (1998). Guidance for industry. Pharmacokinetics in patients with impaired renal function – study design, data analysis and impact on dosing and labelling. Available at URL:http://www.fda.gov/cder/guidance/1449fnl.pdf.

FDA (1999). Draft guidance for industry. Pharmacokinetics in patients with impaired hepatic function:study design, data analysis and impact on dosing and labelling. Available at URL:http://www.fda.gov/cder/guidance/2629dft.htm.

FDA (1999). Guidance for industry. In vivo drug metabolism/drug interaction studies - study design, data analysis, and recommendations for dosing and labelling. Available at URL:http://www.fda.gov/cder/guidance/2635fnl.pdf.

FDA (2000). Guidance for industry. Bioavailability and bioequivalence studies for orally administered drug products – general considerations. Available at URL:http://www.fda.gov/ cder/guidance/3615fnl.pdf.

FDA (2001). Statistical approaches to establishing bioequivalence. Available at URL:http://www.fda.gov/cder/guidance/3616fnl.pdf.

Feng, Z. and Grizzle, J.E. (1992). Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculation. *Statistics in Medicine* **11:**1607-14.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparamatric problems. *The Annals of Statistics* **1:**209-30.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2:**615-29.

Fisher, R.A. (1925). Statistical methods for research workers Oliver and Boyd, Edinburgh, Scotland.

Fisher, R.A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society, Series A* **98:**109-114.

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions, 2nd edition Wiley.

Fleiss, J.L. and Levin, B. (1988). Sample size determination in studies with matched pairs. *Journal of Clinical Epidemiology* **41:**727-730.

Friede, T. and Kieser, M. (2001). A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine* **20:**3861-73.

Friendly, M. (1991). SAS system for statistical graphics. First Edition. Cary, NC, USA.

Frison, L.J., and Pocock, S.J. (1992). Repeated Measures in Clinical Trials:Analysis Using Mean Summary Statistics and Its Implication for Design, *Statistics in Medicine* **11:**1685-1704.

Gail, M.H., Wieand, S. and Piantadosi, S. (1984). Biases estimates of treatment effects in randomised experiments with nonlinear regression and omitted covariates. *Biometrika* **71(3):**431-44.

Garrett, A.D. (2003). Therapeutic equivalence:fallacies and falsification. *Statistics in Medicine* **22:**741-62.

Ghosh, BK (1979). A comparison of some approximate confidence limits for the binomial parameter. Journal of the American Statistical Association 74:894-900.

Gould, A. L. (1992). Interim analyses for monitoring clinical trials that do not materially effect the type I error rate. *Statistics in Medicine* **11:**55-66.

Gould, A. L. (1995). Planning and revising the sample size for a trial. *Statistics in Medicine* **14:**1039-1051.

Gould, A.L. (1995b). Group Sequential Extensions of a Standard Bioequivalence Testing Procedure. *Journal of Pharmacokinetics and Biopharmaceutics* **23:**5–86.

Gould, A.L. (2001). Sample size re-estimation:recent developments and practical considerations. *Statistics in Medicine* **20:**2625-2643.

Gould, A.L. and Shih, W.J (1992). Sample size re-estimation without unblinding for normally distributed data with unknown variance. *Communications in Statistics - Theory and Methods* **21:**2833-53.

Gould, A.L. and Shih, W.J. (1998). Modifying the design of ongoing trials without unblinding. *Statistics in Medicine* **17:**89-100.

Graham, P.L., Mengersen, K. and Morton, A.P. (2003). Confidence limits for the ratio of two rates based on likelihood scores:non iterative method. *Statistics in Medicine* **22:**2071-83.

Greenland, S. (1988). On sample-size and power calculations for studies using confidence intervals. *American Journal of Epidemiology* **128(1),** 231-7.

Greenberg, R.P., Fisher S. (1994). Suspended judgement – seeing through the double masked design:a commentary. *Controlled Clinical Trials* **15:**244-6.

Grieve, A.P. (1989). Confidence intervals and trial sizes. *Lancet* February 11, 337.

Grieve, A.P. (1990). Sample sizes and confidence intervals. *The American Statistician* **44(2),** 190.

Grieve, A.P. (1991). Confidence intervals and sample sizes. *Biometrics* **47:**1597-1603.

Grieve, A.P. (2003). The number needed to treat:a useful clinical measure or a case of the Emperor's new clothes? *Journal of Pharmaceutical Statistics* **2:**87-102.

Guenther, W.C. (1981). Sample size formulas for normal theory t tests. *The American Statistician* **35:**243-4.

Guyatt, G.H., Juniper, E.F., Walter, S.D., Griffith, L.E. and Goldstein, R.S. (1998). Interpreting treatment effects in randomised trials. *British Medical Journal* **316:**690-3.

Hall, P. (1992). Bootstrap and Edgeworth Expansion. Springer-Verlag, New York.

Hamilton, M. (1960). Hamilton Depression Scale. *Journal of Neurology, Neurosurgery and Psychiatry* **23:**56-62.

Hasselblad, V. and Kong, D.F. (2001). Statistical methods for comparison to placebo in active-control trials. *Drug Information Journal* **25:**435-49.

Hauck WW and Anderson S (1992). Types of bioequivalence and related statistical considerations, *International Journal of Clinical Pharmacology, Therapy and Toxicology* **30;**181-187.

Hilton, J.F. and Mehta. C.R. (1993). Power and sample size calculations for exact conditional tests with ordered data. *Biometrics* **49:**609-616.

Hinkley, D. V. (1988). Bootstrap methods. *Journal of the Royal Statistical Society, Series B* **50:**321-337.

Hinkley, D. V. and Schechtman, E. (1987). Conditional bootstrap methods in the mean shift model. *Biometrika* **74:**85-94.

Hudson, D. J. (1966). Fitting segmented curves whose join points have to be estimated. *Journal of the American Statistical Association* **61:**1097-1129.

Hung, H.M.J., Wang, S.J., Lawrence, J. and O'Neil, R.T. (2003). Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine* **22:**213-225.

Hutton, J.L (2000). Number needed to treat:properties and problems (with comments). *Journal of the Royal Statistical Society, Series A* **63(3):**403-419.

Hauck, W.W. and Anderson, S. (1992). Types of bioequivalence and related statistical considerations, *International Journal of Clinical Pharmacology, Therapy and Toxicology* **30:**181-187.

ICH E3 (1996). Structure and content of clinical study reports. Available at URL:http://www.ifpma.org/ich5e.html.

ICH E9 (1998). Statistical principals for clinical trials. Available at URL:http://www.ifpma.org/ich5e.html.

ICH E10 (2000). Choice of control group in clinical trials. Available at URL:http://www.ifpma.org/ich5e.html.

Johnson, W.O., Su, C.L., Gardner, I.A. and Christensen, R. (2004). Sample size calculations for surveys to substantiate freedom of populations from infectious agents. *Biometrics* **60:**165-71.

Johnson, N.L. and Kotz, S. (1994). Distributions in statistics:discrete distributions. John Wiley and Sons, Chichester.

Johnson, N.L. and Kotz, S. (1994). Distributions in statistics:continuous univariate distributions -1. John Wiley and Sons, Chichester (1994).

Johnson, N.L. and Kotz, S. (1994). Distributions in statistics:continuous univariate distributions -2. John Wiley and Sons, Chichester.

Jones, B., Jarvis, P., Lewis, J.A. and Ebbutt, A.F. (1996). Trials to assess equivalence:the importance of rigorous methods. *British Medical Journal* **313:**36-39.

Jones, B and Kenward, M.J. (2003). Design and Analysis of Cross-Over Trials, 2nd ed. Chichester, Wiley.

Jones, D. and Whitehead, J. (1979). Sequential forms of the log rank and modified Wilcoxon tests for censored data. *Biometrika* **66:**105-13

Joseph L, du Berger R, and B´elisle P. (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine*, **16(7):**769-781.

Judson, I, Peiming, M, Peng, B, Verweij, J, Racine, A, Donato di Paulo, E et al (2005). Imatinib pharmacokinetics in patients with gastro-intentistinal stromal tumour: a retrospectice population pharmacokinetics study over time. *Cancer Chemotherapy and Pharmacology* **55:**379-86.

Julious, S.A. (2000). Repeated measures in clinical trials:analysis using means summary statistics and its implications for design. *Statistics in Medicine* **19:**3133-3135.

Julious, S.A. (2001). Inference and estimation in the change point regression problem. *Journal of the Royal Statistical Society, Series D* **50(1):**51-61.

Julious, S.A. (2001). Sample size calculations for early phase trials with uncertain estimates of variability. Conference of Statisticians in the Pharmaceutical Industry, Chester.

Julious, S.A. (2002a). Designing early phase trials with uncertain estimates of variability. International Society for Clinical Biostatisticians Conference , Dijon.

Julious, S.A. (2002b). The number needed to treat:clinically useful measure of treatment effect? Conference of Statisticians in the Pharmaceutical Industry, London.

Julious, S.A. (2004a). Tutorial in Biostatistics:Sample sizes for clinical trials with Normal data. *Statistics in Medicine* **23:**1921-86.

Julious, S.A. (2004b). Designing Clinical Trials with Uncertain Estimates of Variability. *The Journal of Pharmaceutical Statistics* **3:**261-8.

Julious, S.A. (2004c). Using Confidence Intervals Around Individual Means to Assess Statistical Significance between Two Means. *The Journal of Pharmaceutical Statistics* **3:**217-22.

Julious, S.A. (2004d). Sample sizes for non-inferiority studies with binary data. International Society for Clinical Biostatisticians, Leiden.

Julious, S.A. (2004e). Sample size re-determination for repeated measures studies. *Biometrics* **60:**284-5.

Julious, S.A. (2005a). Why do we use pooled variance analysis of variance? *Journal of Pharmaceutical Statistics* **4:**3-5.

Julious, S.A. (2005b). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* (Letter) **24:**3383-4

Julious, S.A. (2005c). Issues with number needed to treat. *Statistics in Medicine* (Letter) **24:**3233-5

Julious, S.A. (2005d). Sample size of twelve per group rule of thumb for a pilot study. *Journal of Pharmaceutical Statistics* **4:**287-91.

Julious, S.A. and Campbell, M.J. (1996). Sample size calculations for ordered categorical data. *Statistics in Medicine* **15:**1065-66.

Julious, S.A. and Campbell, M.J. (1998). Sample sizes for paired or matched ordinal data . *Statistics in Medicine* **17:**1635-1642.

Julious, S.A., Campbell, M.J. and Altman, D.G. (1999). Estimating sample sizes for continuous, binary and ordinal outcomes in paired comparisons:practical hints. *Journal of Biopharmaceutical Statistics* **9(2):**241-251.

Julious, S.A. and Debarnot, C.A.M. (2000). Why are pharmacokinetic data summarised as arithmetic means. *Journal of Biopharmaceutical Statistics* **10(1):**55-71.

Julious, S.A., George, S. and Campbell, M.J. (1995). Sample size for studies using the short form 36 (SF-36). *Journal of Epidemiology and Community Health* **49:**642-644.

Julious, S.A., George, S., Machin, D. and Stephens, R.J. (1997). Sample sizes for randomized trials measuring quality of life in cancer patients. *Quality of Life Research* **6:**109-117.

Julious, S.A. and Khandker, R.K. (2003). Estimating effect sizes for novel health outcomes:stroke impact scale. *DIA Meeting on Patient Outcomes:*Baltimore.

Julious, S.A. and Mullee, M.A. (1994). Confounding and Simpon's paradox. *British Medical Journal* **308**:1480-1.

Julious, S.A. and Mullee, M.A. (2000). Crude rates of outcome. *British Journal of Surgery* **87:**8-9.

Julious, S.A. and Owen, R.J. (2006). Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharmaceutical Statistics* **6(1)**: 29-37.

Julious, S.A. and Patterson, S.D. (2004). Sample sizes for estimation in clinical research. *Journal of Pharmaceutical Statistics* 3:213-5.

Julious, S.A., Walker, S., Campbell, M., George, S.L. and Machin, D. (2000). Determining sample sizes for cancer trials involving quality of life instruments. *British Journal of Cancer* **83(7):**959-963.

Julious, S.A. and Swank, D. (2005). Moving statistics beyond the individual clinical trial- applying decision science to optimise a clinical development plan. *Journal of Pharmaceutical Statistics* **4:**37-46.

Julious, S.A. and Zariffa, N. (2002). The ABC of pharmaceutical trial design:some basic principles. *The Journal of Pharmaceutical Statistic* **1:**45-53.

Keene, O. (2002). Alternatives to the hazard ratio in summarising efficacy in time to event studies:an example from influenza trials. *Statistics in Medicine* **21:**3687-3700.

Kendall, M. and Stuart, A. (1977). The advanced theory of statistics, Volume 1, Distribution Theory, 4th Edition, Charles Griffin & Co, London.

Kieser, M. and Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* **19:**901-911.

Koch, G.G. and Gansky, S.A. (1996). Statistical considerations for multiplicity in confirmatory trials. *Drug Information Journal* **30:**523-534.

Koopman, P.A.R. (1984). Confidence intervals for the ratio of two binomial proportions. *Biometrics* **40:**513-7.

Korn, E.L. (1986) Sample Size Tables for Bounding Small Proportions. *Biometrics* **42:**213-216.

Krams, M., Lees, K.R., Hacke, W., Grieve, A.P. Orgogozo, J.M. and Ford, G.A. (2003). Acute stroke therapy by inhibition in neutrophils (ASTIN). An adaptive dose response study of UK=279,276 in acute ischemic stroke. *Stroke* **34:**2543-48

Krisnaiah, P. K. and Miao, B. Q. (1988). Review about estimation of change-points In:*Handbook of statistics*, Vol 7, Krisnaiah PK and Rao CR (Eds.), North Holland, P375-402.

Kunz, R. and Oxman, A.D. (1998). The unpredictability paradox:review of empirical comparisons of randomised trials and non-randomised clinical trials. *British Medical Journal* **317:**1185-90.

Kupper, L.L. and Hafner, K.B. (1989). How appropriate are popular sample size formulas? *The American Statistician* **43:**101-5.

Kwang, G.P.S. and Hutton, J.L. (2003). Choice of parametric models in survival analysis: applications to monotherapy for epilepsy and cerebral palsy. *Applied Statistics* 2003;**52**(2):153-168.

Lacey, J.M., Keene, O.N. and Pritchard, J.F., and Bye, A. (1997). Common non-compartmental pharmacokinetic variables:are they Normally or log-Normally distributed? *Journal of Biopharmaceutical Statistics* **7(1):**171-178.

Lachin, J.M. (1977). Sample size determination for r x c comparative trials. *Biometrics* **33:**315-24.

Lake, S., Kammann, E., Klar, N., and Betensky, R. (2002). Samples size re-estimation in cluster randomisation trials. *Statistics in Medicine* **21;**1337-50

Laster, L.L. and Johnson, M.F. (2003). Non-inferiority trials:the 'at least as good as' criterion. *Statistics in Medicine* **22:**187-200.

Lee, M.K., Song, H.H., Kang, S.H. and Ahn, C.W. (2002). The determination of sample sizes in the comparison of two multinomial proportions from ordered categories. *Biometrical Journal* **44(4):**395-409.

Lees, B., Molleson, T., Arnett, T. R. and Stevenson, J. C. (1983). Differences in proximal femur bone density over two centuries. *Lancet* **341:**673-5.

Lemeshow, S., Hosmer D.W., Klar, J. and Lwanga, S.K. (1990). Adequacy of sample size in health studies. John Wiley & Sons, Chichester, England.

Lesaffre, E. and Pledger, G. (1999). A note on the number needed to treat. Controlled Clinical Trials **20:**439-47.

Liu, J.P. (1995). Use of the repeated cross-over designs in assessing bioequivalence. *Statistics in Medicine* **14:**1067-78.

Liu, K.J. (1991). Sample sizes for repeated measurements in dichotomous data. *Statistics in Medicine* **10:**463-72.

Lui, K.J. (2001). Interval Estimation of simple difference with dichotomous data with repeated measurements. *Biometrical Journal* **43:**845-61.

Liu, Q., Proschan, M.A. and Pledger, G.W. (2002). A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association* **97:**1034-1041.

Lu, Y. and Bean, J.A. (1995). On the sample size for one-sided equivalence of sensitivities based upon the McNenar's test. *Statistics in Medicine* **14:**1831-9.

Lunn, D.J., Wakefield, J. and Racine-Poon, R. (2001). Cumulatice logic models for ordinal data: a case study involving allergic rhinitis severity scores. *Statistics in Medicine* 2001 **20:**2261-85.

McClung, C. Quessy, S., Julious, S., Segretti, A. and Blum, D. (2004). Placebo response rates in geographical location in COX-2 inhibitor trials of rheumatoid arthritis (RA) and osteoarthritis (OA). American College of Rheumatology Conference, San Antonio.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* **43:**109-142.

Machin, D., Campbell, M.J., Fayers, P., and Pinol, A., (1997). Statistical tables for the design of clinical studies 2nd Ed. Blackwell Scientific Publications, Oxford.

Matthews, J.N.S. (2000). An introduction to randomised controlled trials. Arnold, London.

May, W.L. and Johnson, W.D. (1997). The validity and power for tests of two correlation proportions. *Statistics in Medicine* **16:**1081-96.

Medical Research Council. (1948). Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal* **2:**769-82.

Medical Research Council Lung Cancer Working Party. (1996). Randomised trial of four-drug vs less intensive two-drug chemotherapy in the palliative treatment of patients with small-cell lung cancer (SCLC) and poor prognosis. *British Journal of Cancer* **73:**406-413.

Miettinen, O.S. (1968). The matched pairs design in the case of all-or-none responses. *Biometrics* **24:**339-353.

Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine* **4:**213-226.

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974). *Introduction to the theory of statistics*. London:McGraw-Hill.

Mood, A.M. and Snedecor, G.W. (1946). Query. Biometrics Bulletin **2(6):**120-2.

Moorey, S., Greer, S., Watsonm M., Gormann C., Rowdenn L., Tunmore, R. et al. (1991). The factor structure and factor stability of the Hospital Anxiety and Depression Scale in patients with cancer. *British Journal of Psychiatry* **158:**255-259.

Morikawa, T. and Yoshida, M. (1995). A useful testing strategy in phase III trials:combined test of superiority and test of equivalence. *Journal of Biopharmaceutical Statistics* **5(3):**297-306.

Nam, J. (1997). Establishing equivalence of two treatments and sample size requirements in matched pairs design. *Biometrics* **50:**1422-30.

Newcombe, R.G. (1998a). Two-sided confidence intervals for the single proportion:comparison of seven methods. *Statistics in Medicine* **17:**857-872.

Newcombe, R.G. (1998b). Interval estimation for the difference between independent proportions:comparison of eleven methods. *Statistics in Medicine* **17:**873-890.

Newcombe, R.G. (1998c). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* **17** 2633-2650.

Neyman, J. and Pearson, E.S. (1928). On the use and interpretation of test criteria. *Biometrika* **20(A):**175-294.

Neyman, J. and Pearson, E.S. (1933). On the problem of the most eficient tests of statistical hypotheses. *Philosophical Transitions Royal Society (London)* **231:**289-337.

Neyman, J. and Pearson, E.S. (1933). The testing of statistical hypotheses in relation to the probabilities a priori. *Proceeds of the Cambridge Philosophical Society* **29:**492-510.

Neyman, J, and Pearson, E.S. (1936). Contributions to the theory of testing hypotheses. *Journal Statistical Research Memoirs (University of London)* **1:**1-37.

Neyman, J. and Pearson, E.S. (1936). Sufficient statistics and uniformly most powerful test of statistical hypothesis. *Journal Statistical Research Memoirs (University of London)* **1:**113-137.

Neyman J and Pearson ES (1938). Contributions to the theory of testing statistical hypotheses. *Journal Statistical Research Memoirs (University of London)* **2** 25-57.

NINDS rt-PA Stroke Study Group. (1995). *New England Journal of Medicine* **333:**1581-7.

Nixon, R.M., and Thompson, S.G. (2003). Baseline adjustments for binary data in repeated cross-sectional cluster randomised trials. *Statistics in Medicine* **22:**2673-2692.

Noether, G.E. (1987). Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association* **82:**645-47.

O'Brien, R.G. and Lohr, V.I. (1984). Power analysis for Linear models:the time has come. SUGI Conference proceedings.

O'Quigley J, Pepe M and Fisher L (1990). Continual reassessment method: a practical guide for phase I clinical trials in cancer. *Biometrics* **46:** 33-48.

O'Quigley J and Shen LZ (1996). Continuous reassessment methods a likelihood approach. *Biometrics* **52:** 673-84

Olkin, I. (1998). Odds ratios revisited. *Evidence Based Medicine* **3:**71.

Owen, D.B (1965). A special case of a bivariate non-central t-distribution. *Biometrika* **52:**437-446.

Owen, R (2002). Bayesian approaches to clinical trials. PSI Annual Conference.

Patel, K. Kay, R. and Rowell, L. (2006). Comparing proportional hazards and accelerated failure time models: and application in influenza. *Pharmaceutical Statistics* (In Press).

Pham-Gia, T and Turkkan, N. Sample size determination in Bayesian analysis. The statistician 1992 41:389-92.

Posch, M. and Bauer, P. (2000). Interim analysis and sample size reassessment. *Biometrics* **56:**1170-6.

Proschan, M.A., Liu, Q. and Hunsberger, S. (2003). Practical midcourse sample modification in clinical trials. *Controlled Clinical Trials* **23:**4-15.

Quandt, R.E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association* **53:**873-80.

Quandt, R.E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association* **55:**424-330.

Rabbee, N., Mehta, C., Patel, N. and Senchaudhuri, P. (2003). Power and sample for ordered data. *Statistical Methods in Medical Research* **12:**73-84.

Rao, C.R. (1965). Linear statistical inference and its applications. John Wiley and Sons, Chichester, England.

Rasch, D. and Horrendorfer, G. (1986). Experimental design:sample size determination and block designs. D. Reidel Publishing Company, Lancaster, England.

Reiczigel, J. (2003). Confidence intervals for the binomial parameter:some new considerations. *Statistics in Medicine* **22:**611-21.

Robinson, L.D. and Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* **58(2):**227-40.

Royston, P. (1993). Exact conditional and unconditional sample size for pair-matched studies with binary outcome:a practical guide. *Statistics in Medicine* **12:**699-712.

Schall, R. and Williams, R.L. for the Food and Drug Administration Individual Bioequivalence Working Group (1996). Towards a practical strategy for assessing individual bioequivalence, *Journal of Pharmacokinetics and Biopharmaceutics* **24:**133-149.

Sackett, D.L., Deeks, J.J. and Altman, D.G. (1996). Down with odds-ratios! *Evidence-Based Medicine* **1(6):**164-6.

Santer, T.J. and Snell, M.K. (1980). Small sample confidence intervals for p1-p2 and p1/p2 in 2x2 contingency tables. *Journal of the American Statistical Association* **75:**386-94.

SAS Institute Inc (1985). *SAS/IML User's Guide for Personal Computers*, Version 6 Edition. Cary, NC:SAS Institute Inc.

SAS Institute Inc. (1990). *Language:Reference,* Version 6 First Edition. Cary, NC:SAS Institute Inc.

Schesselman, J.J. (1982). Case-control studies. Oxford University Press, New York.

Schouten, H.J.A. (1999). Sample size formula with a continuous outcome for unequal group sizes and unequal variances. *Statistics in Medicine* **18:**87-91.

Schulzer, M. and Mancini, G.B. (1996). 'Unqualified success' and 'unmitigated failure':number-needed-to-treat-related concepts for assessing treatment efficacy in the presence of treatment-induced adverse events. *International Journal of Epidemiology* **25(4):**704-12.

Seber, G. A. F. (1977) *Linear Regression Analysis.* John Wiley and Sons, Chichester, England.

Senn S. (1993). Cross-over trials in clinical research. John Wiley and Sons, Chichester, England.

Senn, S. (1997). Statistical issues in drug development. John Wiley and Sons, Chichester, England.

Senn, S. (1998). In the blood:proposed new requirements for the registering of generic drugs. *The Lancet* **352:**85-86.

Senn, S. (2000). Consensus and controversy in pharmaceutical statistics (with discussion). *Journal of the Royal Statistical Society, Series D* **49:**135-76.

Senn, S. (2001). Statistical issues in bioequivelence. *Statistics in Medicine* **20:**2787-2799.

Senn, S. (2001). Guest Editorial. The misunderstood placebo. *Applied Clinical Trials* **5:**40-46.

Senn, S, Stevens, L and Chaturvedi, N (2000). Tutorial in Biostatistics: Repeated measures in clinical trials: simple measures for analysis using

Shaban, S.A. (1980). Change-point problem and two-phase regression:an annotated bibliography. *International Statistical Review* **48:**83-93.

Sheiner, L.B. (1997). Learning versus confirming in clinical drug development. *Clinical Pharmacology and Therapeutics* **61:**275-291.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* **2:**238-41.

Singer, J. (2001). A simple procedure to compute sample size needed to compared two independent groups when the population variances are unknown. *Statistics in Medicine* **20:**1089-1095.

Smeeth, L., Haines, A. and Ebrahim, S. (1999). Numbers needed to treat derived from meta analyses – sometimes informative, usually misleading. *British Medical Journal* **318:**1548-1551.

Spiegelhalter, D.J. (2001). Bayesian methods for cluster randomised trials with continuous responses. *Statistical in Medicine* **20:**435:52

Spiegelhalter D., Abrams, K.R., Myles, J.P. (2004). *Bayesian approaches to clinical trials and health-care evaluation.* Wiley: Chichester, England.

Spiegelhalter, D. Freedman, L. and Parmer, M. (1995). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society, Series A* **157:**387-416.

Spiegelhalter, D.J., Myles, J.P., Jones, D.R. and Abrams, K.R. (1999). Methods in health service research: An introduction to Bayesian methods in health service research. *British Medical Journal* **319:**508-12.

Stampfer, M.J., Goldhaber, S.Z., Yusuf, S., Peto, R. and Hennekens, C.H. (1982). Effect of introvenous streptokinase on acute myocardial infarction:pooled results from randomised trials. *New England Journal of Medicine* **307:**1180-82.

Sterne, T.E. (1945). Some Remarks on Confidence or Fiducial Limits. *Biometrika* **41:**275-278.

Swiger, L.A., Harvey, W.R., Everson, D.O. and Gregory, K.E. (1964). The variance of intraclass correlation involving groups with one observation. *Biometrics* **20:**818-26

Tang, M.L. (2003). Matched-pair non-inferiority using rate ratio:a comparison of current methods and sample size refinement. *Controlled Clinical Trials* **24:**364-77.

Tang, N.S., Tang, M.L. and Chan, S.F. (2003). On tests of equivalence via non-unity relative risk for matched pairs design. *Statistics in Medicine* **22:**1217-1233.

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired sample design. *Statistics in Medicine* **17:**891-908.

Tango, T. (1999). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* **18**:3511-3513.

Thompson, S.G., Warn, D.E. and Turner, R.M. (2004). Bayesian methods for analysis of binary data in cluster randomised trials on the absolute risk scale. *Statistics in Medicine* **23:**389-410.

Toendle, J.F. and Frank, J. (2001). Unbiased confidence intervals for the odds ratio of two independent samples with applications to case control data. *Biometrics* **57:**484-89.

Tu, D. (1998). On the use of the ratio or the odds ratio of cure rates in therapeutic equivalence clinical trials with binary endpoints. *Journal of Biopharmaceutical Statistics* **8:**135-76.

Turner, R.M., Omar, R.Z. and Thompson, S.G. (2006). Constructing intervals for the intra-cluster correlation coefficient using Bayesian modelling and application in cluster randomised trials. *Statistics in Medicine* (In Press).

Turner, R.M., Omar, R.Z. and Thompson, S.G. (2001). Bayesian methods of analysis for cluster randomised ttrials with binary outcome data. *Statistics in Medicine* **20:**453-71

Turner, R.M., Prevost, A.T. and Thompson, S.G. (2004). Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randoimised trials. *Statistics in Medicine* **23:**1195-1214

Ukoummunne, O.C. (2003). A comparison of confidence intervals methofs for the intraclass correlation coefficient in cluster randomised trials. *Statistics in Medicine* **21:**3757-74

Ukoummunee, O.C., Davison, A.C., Gulliford, M.C. and Chinn, C. (2003). Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Statistics in Medicine* **22:**3805-21

Vollset, S.E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine* **12**:809-824.

Walker, S. (1998). Odds ratios revisited. Evidence base medicine **3:**71.

Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in Medicine* **12**:2257-72.

Whitehead, A. and Jones, N.M.B. (1994). A meta analysis of clinical trials involving different classifications of response into ordered categories. *Statistics in Medicine* **13:**2503-2515.

Whitehead, A. and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomised trials. *Statistics in Medicine* **10:**1665-77.

Whitehead, J, Zhou Y, Patterson, S, Webber, D and Francis S (2001). Easy to implement Bayesian methods for dose escalation studies in healthy volunteers. *Biostatistics* **2:**47-61.

Wiens, B.L. (2002). Choosing an equivalence limit for noninferiority and or equivalence studies. *Controlled Clinical Trials* **23:**2-14.

Wilson, E.B. (1927). Probably inference, the law of succession and statistical inference. *Journal of the American Statistical Association* **22:**202-12.

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficacy of clinical trials. *Statistics in Medicine* **9:**65-72.

Wood, J. and Lambert, M. (1999). Sample-size calculations for trials in health services research. *Journal of Health Service Research and Policy* **4:**226-229.

Worsley, K.J. (1983). Testing for a two-phase multiple regression. *Technometrics* **25:**35-42.

Wu, C.F.J. (1986). Jackknife, bootstrap and other re-sampling methods in regression analysis. *Annals of Statistics* **14:**1261-1295.

Yoshioka, A. (1998). Use of randomisation in the Medical Research Council's clinical trials of streptomycin in pulmonary tuberculosis in the 1940s. *British Medical Journal* **317**:1220-3.

Zigmond, A.S. and Snaith, R.P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatric Scandinavia* **67:**361-370.

Zucker, D.M. (2004). Sample size re-determination for repeated measures studies. *Biometrics* **60:**284-5.

Zucker, D.M. and Denne, J. (2002). Sample size re-determination for repeated measures studies. *Biometrics* **48(3):**548-59.

Zucker, D.M., Wittes, J.T., Schabenberger, O. and Brittain E. (1999). Internal pilot studies II. Comparison of various procedures. *Statistics in Medicine* **18:**3493-3509.